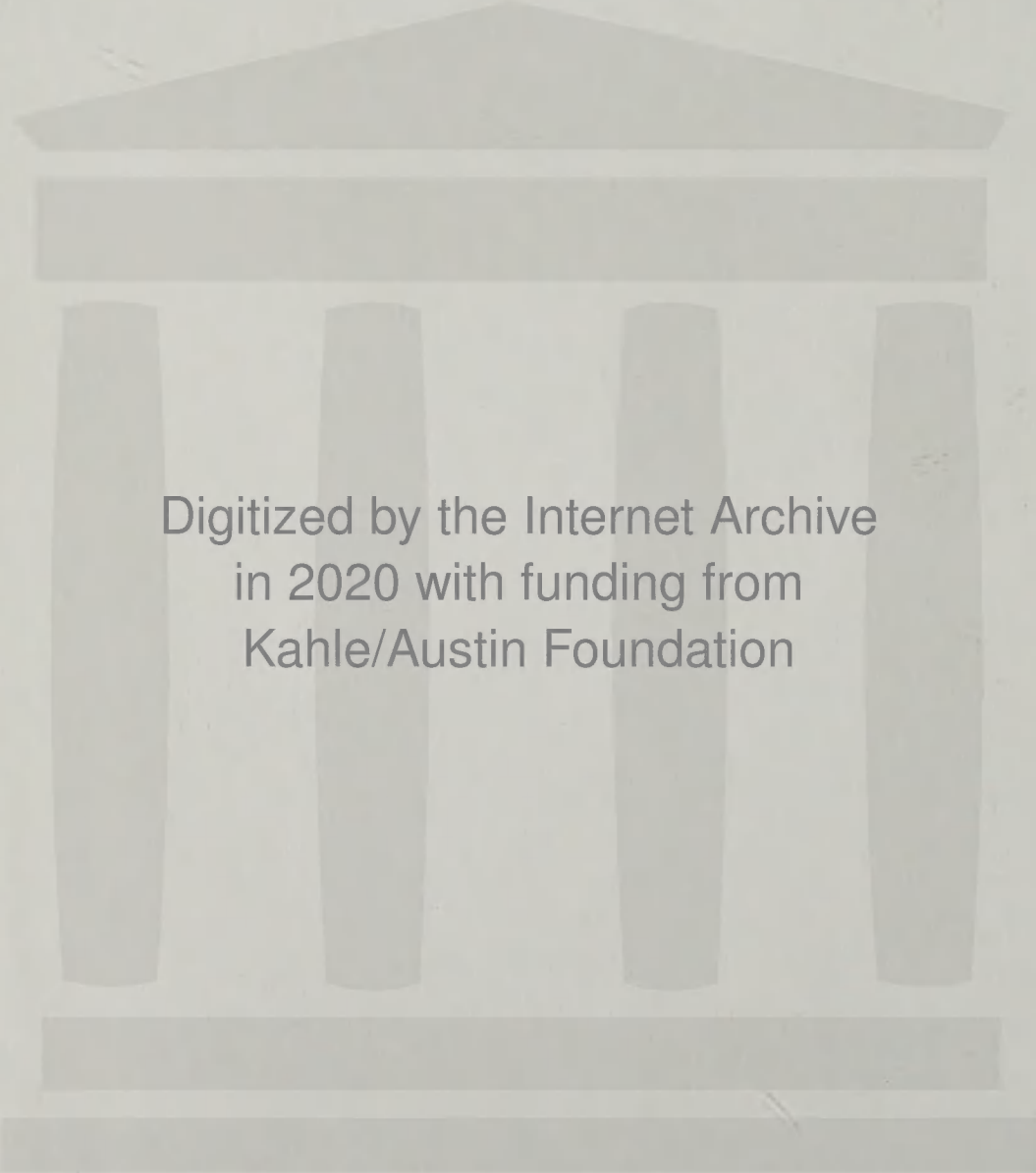


AP 1388

150.5
f 812
V.72



Digitized by the Internet Archive
in 2020 with funding from
Kahle/Austin Foundation

Consulting Editors

Lewis E. Albright, *deRecat & Associates, San Francisco, California*
Earl A. Alluisi, *OUSDRE, The Pentagon, Washington, DC*
Kenneth M. Alvares, *Frito-Lay, Dallas, Texas*
Phipps Arabie, *University of Illinois*
William B. Askren, *Air Force Human Resources Laboratory, Wright-Patterson Air Force Base, Ohio*
Kathryn M. Bartol, *University of Maryland*
Bernard M. Bass, *State University of New York, Binghamton*
Robert S. Billings, *Ohio State University*
Philip Bobko, *University of Kentucky*
C. Alan Boneau, *George Mason University*
Walter C. Borman, *Personnel Decisions Research Institute, Minneapolis, Minnesota*
Donald E. Broadbent, *University of Oxford, England*
Wayne F. Cascio, *University of Colorado, Denver*
Margaret M. Clifford, *University of Iowa*
H. Peter Dachler, *Hochschule St. Gallen für Wirts & Sozialwissen, St. Gallen, Switzerland*
Dan R. Dalton, *Indiana University*
Mark L. Davison, *University of Minnesota*
Robyn M. Dawes, *Carnegie-Mellon University*
Fritz Drasgow, *University of Illinois, Champaign*
Beverly Dugan, *New York Telephone, New York, New York*
E. Ralph Dusek, *JIL Systems and Services, Arlington, Virginia*
James L. Farr, *Pennsylvania State University*
Jack M. Feldman, *Georgia Institute of Technology*
Jeffrey H. Greenhaus, *Drexel University*
Tove Helland Hammer, *Cornell University*

William C. Howell, *Rice University*
Daniel R. Ilgen, *Michigan State University*
Andrew S. Imada, *University of Southern California*
Lawrence R. James, *Georgia Institute of Technology*
Stanislav V. Kasl, *Yale University*
James G. Kelly, *University of Illinois, Chicago*
Gary P. Latham, *University of Washington*
Edwin A. Locke, *University of Maryland*
Robert P. Lowman, *Kansas State University*
Ben B. Morgan, Jr., *University of Central Florida*
Karlene H. Roberts, *University of California, Berkeley*
Paul R. Sackett, *University of Illinois, Chicago*
Steven L. Sauter, *NIOSH, Cincinnati, Ohio*
Frank L. Schmidt, *University of Iowa*
Neal Schmitt, *Michigan State University*
Lyle F. Schoenfeldt, *Texas A&M University*
Stanley E. Seashore, *University of Michigan*
Kirk H. Smith, *Bowling Green State University*
Patricia Cain Smith, *Bowling Green State University*
Barry M. Staw, *University of California, Berkeley*
Mary L. Tenopir, *American Telephone & Telegraph Company, New York, New York*
James R. Terborg, *University of Oregon*
Gary L. Wells, *University of Alberta*
Gary A. Yukl, *State University of New York, Albany*
Sheldon Zedeck, *University of California, Berkeley*

APA Journal Staff

Susan Knapp
Executive Editor

Leslie A. Cameron
Director, Journals Program

Lois Czapiewski and Theodore J. Baroody
Production Editors

W. Ralph Eubanks
Manager, Journal Production

Jodi Ashcraft
Advertising Sales Manager

Published quarterly
by the
American Psychological
Association

Journal of
Applied
Psychology

Editor

Robert M. Guion

Associate Editors

Irwin L. Goldstein

Frank J. Landy

The *Journal of Applied Psychology* is devoted primarily to original investigations that contribute new knowledge and understanding to any field of applied psychology except clinical psychology. The journal considers quantitative investigations of interest to psychologists doing research or working in such settings as universities, industry, government, urban affairs, police and correctional systems, health and educational institutions, transportation and defense systems, and consumer affairs. A theoretical or review article may be accepted if it represents a special contribution to an applied field.

Editor

Robert M. Guion, *Bowling Green State University*

Associate Editors

Irwin L. Goldstein, *University of Maryland*

Frank J. Landy, *Pennsylvania State University*

Consulting Editors

Lewis E. Albright, *deRecat & Associates, San Francisco, California*
Earl A. Alluisi, *Air Force Human Resources Laboratory, Brooks Air Force Base, Texas*

Kenneth M. Alvares, *Frito-Lay, Dallas, Texas*

Phipps Arabie, *University of Illinois*

William B. Askren, *Universal Energy Systems, Dayton, Ohio*

Kathryn M. Bartol, *University of Maryland*

Bernard M. Bass, *State University of New York, Binghamton*

Robert S. Billings, *Ohio State University*

Philip Bobko, *University of Kentucky*

C. Alan Bonneau, *George Mason University*

Walter C. Borman, *Personnel Decisions Research Institute, Minneapolis, Minnesota*

Donald E. Broadbent, *University of Oxford, England*

Wayne F. Cascio, *University of Colorado, Denver*

Margaret M. Clifford, *University of Iowa*

H. Peter Dachler, *Hochschule St. Gallen für Wirts & Sozialwissen, St. Gallen, Switzerland*

Dan R. Dalton, *Indiana University*

Mark L. Davison, *University of Minnesota*

Robyn M. Dawes, *Carnegie-Mellon University*

Fritz Drasgow, *University of Illinois*

Beverly Dugan, *New York Telephone, New York, New York*

E. Ralph Dusek, *Advanced Resource Development Corporation, Columbia, Maryland*

James L. Farr, *Pennsylvania State University*

Jack M. Feldman, *University of Texas, Arlington*

Jeffrey H. Greenhaus, *Drexel University*

Tove Helland Hammer, *Cornell University*

William C. Howell, *Rice University*

Daniel R. Ilgen, *Michigan State University*

Andrew S. Imada, *University of Southern California*

Lawrence R. James, *Georgia Institute of Technology*

Stanislav V. Kasl, *Yale University*

James G. Kelly, *University of Illinois*

Gary P. Latham, *University of Washington*

Edwin A. Locke, *University of Maryland*

Robert P. Lowman, *Kansas State University*

Ben B. Morgan, Jr., *Old Dominion University*

Karlene H. Roberts, *University of California, Berkeley*

Paul R. Sackett, *University of Illinois, Chicago*

Steven L. Sauter, *NIOSH, Cincinnati*

Frank L. Schmidt, *University of Iowa*

Neal Schmitt, *Michigan State University*

Lyle F. Schoenfeldt, *Texas A&M University*

Stanley E. Seashore, *University of Michigan*

Kirk H. Smith, *Bowling Green State University*

Patricia Cain Smith, *Bowling Green State University*

Barry M. Staw, *University of California, Berkeley*

Mary L. Tenopir, *American Telephone & Telegraph Company, New York*

James R. Terborg, *University of Oregon*

Gary L. Wells, *University of Alberta*

Gary A. Yukl, *State University of New York, Albany*

Sheldon Zedeck, *University of California, Berkeley*

Manuscripts: Submit manuscripts in quadruplicate to the Editor, Robert Guion, Department of Psychology, Bowling Green State University, Bowling Green, OH 43403, according to instructions elsewhere in this journal (see the table of contents). APA and the editors assume no responsibility for statements and opinions advanced by contributors to *Journal of Applied Psychology*.

Change of Address: Send change of address notice and a recent mailing label to the attention of the Subscription Section, APA, 30 days prior to the actual change of address. APA will not replace undelivered copies resulting from address changes; journals will be forwarded only if subscribers notify the local post office in writing that they will guarantee second-class forwarding postage.

Reprints: Authors may order reprints of their articles from the printer when they receive proofs.

Single Issues, Back Issues, and Back Volumes: For information regarding single issues, back issues, or back volumes write to Order Department, APA, 1200 Seventeenth Street, N.W., Washington, DC 20036.

Microform Editions: For information regarding microform editions write to either of the following: University Microfilms, Ann Arbor, MI 48106; or Princeton Microfilms, Princeton, NJ 08540.

Copyright and Permission: Authors must secure from APA and the author of reproduced material written permission to reproduce an article in full or text of more than 500 words. APA normally grants permission contingent upon like permission of the author, inclusion of the APA copyright notice on the first page of reproduced material, and payment of a fee of \$20 per page. Permission from APA and fees are waived for authors who wish to reproduce a single table or figure provided the author's permission is obtained and full credit is given to APA as copyright holder and to the author through a complete citation. Permission and fees are waived for authors who wish to reproduce their own material for personal use; fees only are waived for authors who wish to use more than a single table or figure of their own material commercially. Permission and fees are waived for the photocopying of isolated articles for nonprofit classroom or library reserve use by instructors and educational institutions. Access services may use abstracts without the permission of APA or the author. Libraries are permitted to photocopy beyond the limits of U.S. copyright law; (a) post-1977 articles, provided the per-copy fee in the code for this journal (0021-9010/87/\$00.75) is paid through the Copyright Clearance Center, 27 Congress Street, Salem, MA 01970; (b) pre-1978 articles, provided the per-copy fee stated in the Publishers' Fee List is paid through the Copyright Clearance Center, 27 Congress Street, Salem, MA 01970. Address requests for reprint permission to the Permissions Office, APA, 1200 Seventeenth Street N.W., Washington, DC 20036.

APA Journal Staff: Susan Knapp, *Executive Editor*; Leslie A. Cameron, *Director, Journals Program*; W. Ralph Eubanks, *Manager, Journal Production*; Lois Czapiewski, *Production Editor*; Jodi Ashcraft, *Advertising Sales Manager*.

The *Journal of Applied Psychology* (ISSN 0021-9010) is published quarterly (beginning in February) in one volume per year by the American Psychological Association, Inc., 1400 North Uhle Street, Arlington, VA 22201. Subscriptions are available on a calendar year basis only (January through December). The 1987 rates follow: *Non-member Individual*: \$40 Domestic, \$43 Foreign, \$50 Air Mail. *Institutional*: \$82 Domestic, \$89 Foreign, \$96 Air Mail. *APA Member*: \$30. Printed in the U.S.A. Second-class postage paid at Arlington, VA, and at additional mailing offices. POSTMASTER: Send address changes to the *Journal of Applied Psychology*, 1400 North Uhle Street, Arlington, VA 22201.

-
- 3 Meta-Analysis for Integrating Study Outcomes: A Monte Carlo Study of Its Susceptibility to Type I and Type II Errors
Paul E. Spector and Edward L. Levine
- 10 A Decision-Theoretic Approach to the Use of Appropriateness Measurement for Detecting Invalid Test and Scale Scores
Fritz Drasgow and Elaine Guertler
- 19 Study of the Measurement Bias of Two Standardized Psychological Tests
Fritz Drasgow
- 30 The Power of the Schmidt and Hunter Additive Model of Validity Generalization
Edward R. Kemery, Kevin W. Mossholder, and Lawrence Roth
- 38 Test of the Cultural Bias Hypothesis: Some Israeli Findings
Moshe Zeidner
- 49 Comparison of Several Procedures for Generating J-Coefficients
John W. Hamilton and Terry L. Dickinson
- 55 Reactions to Procedural Injustice in Payment Distributions: Do the Means Justify the Ends?
Jerald Greenberg
- 62 Intentionally Favored, Unintentionally Harmed? Impact of Sex-Based Preferential Selection on Self-Perceptions and Self-Evaluations
Madeline E. Heilman, Michael C. Simon, and David P. Repper
- 69 A Revision of the Job Diagnostic Survey: Elimination of a Measurement Artifact
Jacqueline R. Idaszak and Fritz Drasgow
- 75 Employee Reactions to Workspace Characteristics
Greg R. Oldham and Yitzhak Fried
- 81 Member Variation, Recognition of Expertise, and Group Performance
Robert Libby, Ken T. Trotman, and Ian Zimmer
- 88 Routinization of Mental Training in Organizations: Effects on Performance and Well-Being
Gerry Larsson
- 97 Effects of Categorization, Attribution, and Encoding Processes on Leadership Perceptions
Steven F. Cronshaw and Robert G. Lord
- 107 Task Planning and Energy Expended: Exploration of How Goals Influence Performance
P. Christopher Earley, Pauline Wojnaroski, and William Prest
- 115 Patterns of Work and Nonwork Satisfaction
Garnett Stokes Shaffer

- 125 Experimental Test of an Emotion-Based Approach to Fitting Brand Names to Products
Albert Mehrabian and Robert de Wetter
- 131 Effects of Negative Ions on Cognitive Performance
Robert A. Baron

- 138 Receptivity and Planned Change: Community Attitudes and Deinstitutionalization
Gregory H. Wilmoth, Starr Silver, and Lawrence J. Severy

Short Notes

- 146 Averaging Correlation Coefficients: Should Fisher's z Transformation Be Used?
N. Clayton Silver and William P. Dunlap
- 149 Reexamining the Component Stability of Owens's Biographical Questionnaire
Gary J. Lautenschlager and Garnett Stokes Shaffer
- 153 Modification of the Minnesota Clerical Test to Predict Performance on Video Display Terminals
Edward M. Silver and Corwin Bennett
- 156 Sex Effects in Workplace Justice Outcomes: A Field Assessment
Dan R. Dalton, William D. Todor, and Crystal L. Owen
- 160 Predictive Validity of the MODE Conflict Instrument
Boris Kabanoff

Monograph

- 165 A Butterfly Catastrophe Model of Motivation in Organizations: Academic Performance
Stephen J. Guastello

Other

- 137 Instructions to Authors
- 54 Low Publication Prices for APA Members and Affiliates
- 124 Searches Open for Editors of Five APA Journals

Meta-Analysis for Integrating Study Outcomes: A Monte Carlo Study of Its Susceptibility to Type I and Type II Errors

Paul E. Spector and Edward L. Levine
University of South Florida

A Monte Carlo study was conducted to determine Types I and II error rates of the Schmidt and Hunter (S&H) meta-analysis method and the U statistic for assessing homogeneity within a set of correlations. One thousand samples of correlations were generated randomly to fill each of 450 cells of an $18 \times 5 \times 5$ (Underlying Population Correlations \times Numbers of Correlations Compared \times Sample Size Per Correlation) design. To assess Type I error rates, correlations were drawn from the same population. To assess power, correlations were drawn from two different populations. As compared with U , which was uniformly robust, the Type I error rate for the S&H method was unacceptably high in many cells, particularly when the criterion for determining homogeneity was set at a highly conservative level. Power for the S&H method increased with increasing size of population differences, sample size per correlation, and in some cases, number of correlations compared. The U statistic did more poorly in most conditions in protecting from Type II errors.

Cogent arguments have recently been advanced in favor of meta-analyses, quantitative methods for integrating findings across studies (Glass, 1976; Hunter, Schmidt, & Jackson, 1982; Rosenthal, 1984). Hunter et al. (1982) advocated the use of methods they developed to aggregate strength of association results statistically across studies.

The Hunter et al. (1982) approach consists of several analytical steps. First, a particular descriptive statistic to measure degree of relation, such as the correlation coefficient, is chosen. A thorough literature review is conducted, and the statistics of choice are extracted from each study deemed appropriate for inclusion in the data base. The set of statistics becomes the data for the meta-analysis. The analysis begins by calculating the mean of the statistics. Next, the variance of the effect size statistics is calculated. For correlation coefficients, the population formula is given by Hunter et al. (1982, p. 41) as

$$s_r^2 = \Sigma[N(r - \bar{r})^2] / \Sigma N,$$

where r = measure of effect size and N = sample size per r .

The variance is corrected for statistical artifacts, such as sampling error, differential reliability, and restriction of range. This corrected variance is used to fit confidence intervals around the mean effect size to determine if outcomes across studies have been positive. It is also compared to the uncorrected variance to determine if additional moderator variables might be implicated in accounting for variability in sample statistics across studies. A moderator variable in this context refers to some situational or personal characteristic that is associated with differences in study outcomes. For example, in a test of job character-

istics theory, Spector (1985) found that the relation between job scope and employee outcomes was affected by the level of higher order need strength of those employees. In a variant of meta-analysis that Hunter et al. (1982) referred to as the Schmidt and Hunter (S&H) procedure, the criterion for concluding that more than artifacts is responsible for variance is that the variance attributable to such artifacts as sampling error across studies is less than 75% of the observed variance in study outcomes.

When artifacts fail to account for 75% of the variance, a search for moderator variables is indicated. Moderators will manifest themselves by correlating strongly with study outcomes, or by allowing for a categorization of studies into clusters that in themselves display differences in average outcomes and less variance among effect size statistics than the original entire sample.

The use of meta-analysis, although a recent development, has increased exponentially over the past 7 years, and has found particular application in industrial and organizational psychology (Jackson, 1984). Meta-analyses have apparently resulted in the resolution of several critically important questions. For example, Hunter and Hunter (1984) were able (a) to conclude that cognitive ability tests predict job performance, and (b) to estimate productivity gains in dollars accruing to organizations that use such test scores to rank job applicants.

Despite its increasingly widespread use, an overriding issue that remains to be decided in the use of meta-analysis is how to compute the variance in the effect sizes or correlation coefficients across studies (e.g., Raju & Burke, 1983). On one hand, an erroneous variance estimate increases the probability that conclusions about effect size in the population will likewise be erroneous. For example, a variance estimate for validity coefficients that is too large would produce a confidence interval around the observed mean that would be too likely to include a value of zero, or near zero. On the other hand, a faulty variance estimate could lead to erroneous conclusions about the presence or absence of moderators. An estimate of variance that is

The contributions of both authors were equal. Order of authorship was determined by a coin flip.

Correspondence concerning this article should be addressed to Paul E. Spector, Department of Psychology, University of South Florida, Tampa, Florida 33620.

too small fails to suggest the presence of moderators unduly often. An estimate that is too large suggests the presence of moderators too often.

Hunter et al. (1982) advanced the notion that the S&H approach to the estimation of the variance of effect sizes gives a better approximation to its true value than that of Glass (1976). By contrast with Glass, who relies on the average variance across studies, Hunter et al. corrected (i.e., reduced) the observed variance in effect sizes across studies for statistical artifacts, as previously discussed. The primary artifact in most applications of the S&H technique is the variance in observed correlations due to sampling error. This approach has been used in a number of studies by Hunter, Schmidt, and their colleagues, and has resulted in the startling finding that validities of cognitive ability tests generalize across—or in other words are not moderated by—a variety of situational constraints like organizational setting and job type (e.g., Pearlman, Schmidt, & Hunter, 1980; Schmidt, Hunter, & Pearlman, 1981). This finding is startling because it has been an almost axiomatic belief in the industrial and organizational field that test validity is situation specific; that is, validity of a given test varies as a function of organizational, job, and applicant characteristics.

Given that the Schmidt and Hunter procedure is relatively new and controversial, and that some results with the procedure are counterintuitive in failing to find moderator effects, we wondered whether it has sufficient statistical power to detect a moderator effect when it in fact exists. In other words, how much statistical power does the technique have under a variety of conditions to protect from making a Type II error? This led us immediately to the complementary question of the susceptibility of the technique to Type I errors. We were specifically concerned with the behavior of the technique under conditions in which the only artifact was sampling error, which generally accounts for the major portion of corrections in observed variance with the Schmidt and Hunter procedure, relative to other artifacts such as reliability of criteria. These problems have also been studied with Monte Carlo techniques by Osburn, Calender, Greener, and Ashworth (1983). The current study will both replicate and extend their findings in a variety of ways.

To test the susceptibility of the Schmidt and Hunter procedure to Type I and II errors, we conducted a Monte Carlo study using the Pearson product-moment correlation as the measure of effect size. Multiple samples of correlations were generated from normal populations with specified parameters (mean and standard error). Varied were the number of correlations aggregated and the underlying sample size (number of subjects) per correlation. For each combination of population values, number of correlations, and sample sizes, 1,000 replications were generated, each subject to a meta-analysis. The results of the analyses were aggregated across the 1,000 replications. The current study extends the Osburn et al. (1983) results by including a wider range of sample sizes and numbers of correlations compared.

For each of the replications the variance among the sample correlations was calculated (see the Method section for the actual formulas used). Because the only artifact was sampling error, this was calculated according to the S&H procedure. We chose to deal only with sampling error for two reasons. First, in many applications of this technique, only corrections for sam-

pling error are reported, primarily because data necessary for other corrections, such as reliability or restriction of range, are unavailable in studies being cumulated. Second, sampling error typically accounts for a significant majority of artifact variance when the procedure is actually used. Furthermore, the focus on a single artifact, with data meeting all distributional assumptions, will indicate how well the procedure works under the best conditions. If the procedure is found to perform acceptably, further work with additional artifact adjustments and violation of distributional assumptions would seem in order.

Once the sampling error variance is calculated, the ratio of error to sample variance is derived. According to one variant of the S&H procedure, if this ratio is less than .75, one would conclude that more than error variance is accounting for variation among correlations in the sample. With data of the sort used here, a case may be made for using a 95% rather than 75% criterion (cf. Osburn et al., 1983). Therefore, both the 75% (S&H-75) and 95% (S&H-95) procedures were compared. Because the only artifact in these data was sampling error, an alternative means of detecting moderators would be a test for homogeneity within a set of correlations. Such a statistic is *U* (Marascuilo, 1971), which is a statistical test for determining if at least one of a set of correlations is significantly different from the others. It is distributed approximately as chi-square. Thus, we compared three procedures for detecting moderators within a set of correlations, S&H-75, S&H-95, and *U*. The inclusion of these last two decision criteria is another extension to the Osburn et al. (1983) study.

The Type I error rate can be calculated by aggregating the number of times the S&H procedure and *U* statistic detect differences among correlations when all are from the same population. This would be represented by the S&H procedure, failing to find that 75% (or 95%) of the variance among correlations is accounted for by sampling error or the *U* being statistically significant. The Type II error rate can be calculated by aggregating the same outcomes in situations in which the correlations come from two (or more) populations with different means. Finally, one additional indication of bias for the S&H procedure would be detected by calculating across the 1,000 replications, the mean of the ratios of error to sample variance when all correlations are from the same population. In this case, all sample variance is error variance, and therefore, the expected value of the ratio should be 1.0. Deviation from this ratio would suggest a bias in the S&H method of calculating sample variance. Although Hunter et al. (1982, p. 142) were aware of the problems of Type I and Type II errors in meta-analysis, they did not appear to view these errors as a result of bias in their statistical estimations of variance of outcomes across studies.

Method

Design

Although comparisons among conditions were not made in this study, it can be conceptualized as an $18 \times 5 \times 5$ completely crossed design, with variables, population correlation, number of correlations compared, and sample size per correlation. Eighteen different conditions of population correlations were chosen. For 3 of them, all sample correlations came from populations with identical means (0, .3, or .5). For the remaining 15 conditions, one half of the correlations came from

Table 1
Type I Error Rates for Schmidt and Hunter (S&H) Method and U Statistic

No. of rs	Sample size per r														
	S&H-75					S&H-95					U				
	30	75	100	250	500	30	75	100	250	500	30	75	100	250	500
Population value = 0															
6	26.6	25.4	24.7	23.0	25.4	41.8	39.6	41.1	38.8	37.2	5.5	4.8	4.0	5.1	4.4
10	23.1	22.3	22.3	20.0	21.2	41.3	41.0	41.4	42.7	39.8	5.3	4.7	5.8	6.5	5.6
20	16.9	15.7	15.8	15.0	15.3	45.2	39.9	43.7	39.8	39.8	5.7	5.0	5.3	5.1	5.9
40	10.2	10.0	8.1	7.5	8.0	46.6	39.0	39.9	41.4	37.1	4.9	4.2	5.5	4.8	6.8
100	2.6	1.6	2.0	1.5	1.4	43.0	38.6	37.6	35.4	36.2	5.4	5.9	4.3	4.3	5.4
Population value = .3															
6	28.0	26.2	27.7	25.1	24.1	43.0	38.6	39.7	38.7	38.2	4.8	4.7	6.6	5.5	5.4
10	22.9	20.9	22.7	19.2	24.4	46.9	41.0	43.1	41.1	39.4	5.5	4.4	4.9	5.1	5.6
20	17.5	15.2	15.0	16.3	14.5	46.0	42.1	42.2	40.7	41.1	5.8	6.1	5.4	5.1	4.5
40	13.3	7.9	9.2	8.1	8.1	46.1	41.1	38.5	39.0	39.7	4.6	4.1	5.4	5.2	5.5
100	3.3	1.9	2.2	1.6	1.4	45.7	39.7	37.6	34.7	36.0	4.2	5.1	5.0	4.4	3.9
Population value = .5															
6	32.2	26.7	26.4	26.3	26.4	44.4	38.5	41.8	41.8	39.7	4.3	5.9	5.7	5.2	4.7
10	25.9	22.1	21.8	19.3	22.0	47.0	41.7	41.9	38.6	38.4	5.9	5.0	4.3	4.1	5.6
20	20.1	16.7	14.4	13.8	15.2	43.9	40.2	43.5	40.0	37.2	4.3	5.4	5.3	5.9	6.7
40	12.2	8.3	9.9	7.3	9.2	48.2	43.2	42.6	39.4	37.9	4.2	5.1	4.3	5.3	6.1
100	4.0	2.8	1.7	1.8	2.0	51.9	45.0	37.5	36.4	36.8	5.7	5.8	5.0	5.6	4.7

populations with one mean, and one half from populations with another. The specific comparisons are shown in Table 2 (on p. 6) and were all combinations of 0, .1, .2, .3, .4, and .5. These comparisons were chosen to represent the types of differences likely to occur as the result of moderators in psychological research. Five levels of number of correlations were chosen: 6, 10, 20, 40, and 100. Likewise, five levels of sample size per correlation were used: 30, 75, 100, 250, and 500. These were chosen to represent the range of sample sizes likely to be found in meta-analyses. For every individual meta-analysis conducted, sample sizes were kept constant across all correlations. Although this restriction is unnecessary in actual meta-analyses, it was done here for convenience. Formulas will be shown with this simplifying restriction.

Data Sets

Data for this study were correlations generated from normal distributions with selected population means and standard deviations. To produce normal distributions of correlations, Fisher's z values were generated and transformed to corresponding rs by reversing the r to z transformation formula. Means were chosen to fill out the experimental design of the study, representing the 18 levels previously described. Standard deviations of the z transformed rs were based on the error variance formula

$$\sigma_e = [1/(n - 3)]^{1/2}.$$

All samples were generated using the SAS (Statistical Analysis System; SAS Institute, 1982) random number generator for normally distributed data (RANNOR). For each cell of the design, 1,000 samples (replications) of correlations were generated. The actual number of correlations per replication was determined in accordance with the particular condition in the overall study design.

Procedure

For each of the 18 × 5 × 5 conditions of this study, 1,000 samples of correlations were generated. For each of the 1,000, the S&H-75, S&H-95, and the U statistic were applied. The S&H procedures required the calculation of variance due to sampling error using the formula

$$s_e^2 = (1 - \bar{r}^2)^2/n,$$

where \bar{r} = mean of sample correlations, and n = sample size per correlation (Hunter et al., 1982, p. 44). The variance among the sample correlations was calculated using the formula

$$s_r^2 = \Sigma(r - \bar{r})^2/(nc - 1),$$

where r = observed correlation, and nc = number of correlations (this formula differs from Hunter et al., 1982, p. 41, in using $n - 1$ rather than n in the denominator). Both of these formulas have been simplified for the special case of equal sample sizes for all correlations.

The S&H criterion for concluding that a group of correlations come from at least two populations is that the predicted error variance (s_e^2) is less than 75% (or 95%) of the observed variance (s_r^2). For these determinations, the ratio of error to correlation variance (s_e^2/s_r^2) was calculated for each analysis conducted. The U statistic was calculated with the formula

$$U = (n - 3)\Sigma(z - \bar{z})^2,$$

where z = z transformed r, and \bar{z} = mean of zs. It is distributed as chi-square with $nc - 1$ degrees of freedom, where nc = number of correlations. A significant U indicates that a group of correlations comes from at least two populations.

For each of the 450 conditions of this study, results were summarized across the 1,000 replications. The percentage of times each of the three

Table 2
Power of Schmidt and Hunter (S&H) Method and U Statistic

No. of rs	Sample size per r														
	S&H-75					S&H-95					U				
	30	75	100	250	500	30	75	100	250	500	30	75	100	250	500
Population value = .5, .4															
6	33.3	44.7	47.8	73.8	93.3	49.4	56.2	60.4	84.2	95.9	8.0	12.4	16.8	42.3	74.1
10	33.4	46.7	55.3	79.1	96.5	54.8	62.7	69.0	91.0	98.6	9.9	16.7	22.7	54.4	88.5
20	29.7	45.4	52.9	89.4	99.7	59.0	73.2	81.0	97.8	100	10.3	22.2	31.9	75.5	98.9
40	24.6	43.6	56.6	96.5	99.9	63.0	84.3	89.4	99.8	100	10.6	34.0	47.8	94.0	100
100	15.3	42.9	62.4	99.9	100	76.8	92.3	97.8	100	100	19.6	54.8	77.7	100	100
Population value = .5, .3															
6	51.3	75.2	85.3	99.4	100	65.4	88.6	93.0	100	100	16.1	47.5	60.2	95.9	100
10	55.4	84.9	91.7	99.9	100	73.1	93.4	97.1	99.9	100	21.6	57.1	73.8	99.2	100
20	61.2	92.9	96.8	100	100	81.5	98.6	99.8	100	100	31.1	79.9	92.0	100	100
40	61.2	97.6	100	100	100	90.4	99.9	100	100	100	42.4	96.2	99.8	100	100
100	69.6	100	100	100	100	98.5	100	100	100	100	75.8	100	100	100	100
Population value = .5, .2															
6	70.9	95.8	99.2	100	100	81.0	97.4	99.3	100	100	35.2	78.5	92.6	100	100
10	76.9	98.8	99.7	100	100	89.5	99.5	100	100	100	47.4	93.4	98.5	100	100
20	86.3	99.9	100	100	100	97.3	100	100	100	100	71.5	99.6	99.9	100	100
40	95.8	100	100	100	100	99.3	100	100	100	100	88.0	100	100	100	100
100	99.5	100	100	100	100	100	100	100	100	100	99.9	100	100	100	100
Population value = .5, .1															
6	84.9	99.6	100	100	100	92.4	99.8	100	100	100	56.5	96.9	99.3	100	100
10	92.7	100	100	100	100	98.4	100	100	100	100	71.8	99.8	100	100	100
20	98.3	100	100	100	100	99.7	100	100	100	100	92.9	100	100	100	100
40	100	100	100	100	100	99.9	100	100	100	100	99.5	100	100	100	100
100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Population value = .5, 0															
6	94.0	100	100	100	100	97.7	100	100	100	100	77.2	99.8	100	100	100
10	98.9	100	100	100	100	99.7	100	100	100	100	93.0	100	100	100	100
20	100	100	100	100	100	100	100	100	100	100	99.5	100	100	100	100
40	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Population value = .4, .3															
6	31.7	38.4	44.9	70.6	89.0	47.1	57.2	58.7	78.1	94.3	6.5	12.8	14.6	33.0	67.7
10	31.8	40.8	45.7	74.6	95.4	51.6	59.5	67.3	86.8	99.2	8.2	15.4	17.9	46.1	83.7
20	28.2	38.2	44.7	83.1	99.1	57.1	66.4	74.1	96.0	100	11.8	17.0	25.6	68.6	97.2
40	21.6	38.8	47.4	91.2	99.9	60.3	73.6	85.1	99.3	100	12.7	23.6	39.0	87.9	100
100	10.8	32.2	44.5	99.2	100	72.5	89.5	94.9	100	100	15.9	45.6	62.7	99.9	100
Population value = .4, .2															
6	44.9	70.7	83.3	98.4	100	62.7	79.9	89.4	99.7	100	15.4	37.8	49.5	93.2	100
10	51.3	79.0	86.8	99.7	100	68.7	88.2	94.7	100	100	17.6	45.4	65.2	98.4	100
20	50.6	85.6	96.0	100	100	80.7	95.1	99.1	100	100	26.1	67.1	86.3	100	100
40	55.1	94.5	98.9	100	100	86.3	99.7	99.8	100	100	39.0	91.9	98.0	100	100
100	55.3	99.6	100	100	100	97.1	100	100	100	100	68.5	99.9	100	100	100
Population value = .4, .1															
6	67.9	93.0	97.2	100	100	80.9	97.1	99.2	100	100	30.9	77.5	88.9	100	100
10	72.7	97.7	99.6	100	100	86.1	99.1	99.9	100	100	39.5	88.9	96.2	100	100
20	80.8	100	100	100	100	94.6	100	100	100	100	60.2	98.2	99.9	100	100
40	90.6	100	100	100	100	98.7	100	100	100	100	82.2	100	100	100	100
100	97.0	100	100	100	100	100	100	100	100	100	98.9	100	100	100	100

Table 2 (Continued)

No. of rs	Sample size per r														
	S&H-75					S&H-95					U				
	30	75	100	250	500	30	75	100	250	500	30	75	100	250	500
Population value = .4, 0															
6	83.7	99.3	100	100	100	89.0	99.6	100	100	100	52.9	93.6	98.8	100	100
10	90.1	100	100	100	100	95.7	100	100	100	100	65.3	99.5	100	100	100
20	96.2	100	100	100	100	99.1	100	100	100	100	87.8	100	100	100	100
40	99.7	100	100	100	100	100	100	100	100	100	98.9	100	100	100	100
100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Population value = .3, .2															
6	30.4	39.8	42.9	65.3	87.0	47.4	55.4	57.6	74.2	91.9	7.2	10.8	13.6	30.3	61.8
10	32.2	38.1	41.1	71.3	90.9	50.0	61.0	63.2	84.3	97.4	5.8	13.4	17.3	39.3	76.4
20	26.6	35.1	39.7	78.9	97.7	55.1	62.7	74.2	92.7	99.4	7.2	15.6	21.4	60.2	93.7
40	17.8	32.8	41.0	86.3	99.7	57.7	72.7	80.0	98.3	100	9.4	23.5	31.3	83.0	99.3
100	9.0	24.4	39.9	97.3	100	66.3	83.7	92.2	100	100	14.5	40.0	52.5	98.6	100
Population value = .3, .1															
6	45.5	69.2	78.2	99.0	100	58.7	82.5	87.8	99.1	100	13.6	34.5	48.2	90.2	99.7
10	48.2	74.9	86.2	99.6	100	67.5	84.4	91.5	100	100	18.1	45.2	58.1	98.4	100
20	45.8	84.2	93.8	100	100	76.1	95.6	97.7	100	100	20.8	65.4	81.9	100	100
40	48.8	92.9	98.3	100	100	86.0	99.2	100	100	100	35.5	86.4	95.8	100	100
100	49.8	98.3	100	100	100	95.4	100	100	100	100	58.0	99.7	100	100	100
Population value = .3, 0															
6	64.6	90.2	96.3	100	100	76.0	94.9	98.1	100	100	27.9	68.7	84.6	99.9	100
10	70.8	95.5	99.2	100	100	84.0	98.7	99.2	100	100	36.1	84.0	94.7	100	100
20	77.4	99.9	100	100	100	92.1	99.9	100	100	100	52.0	97.0	99.7	100	100
40	84.6	100	100	100	100	98.8	100	100	100	100	79.4	99.9	100	100	100
100	96.3	100	100	100	100	100	100	100	100	100	97.5	100	100	100	100
Population value = .2, .1															
6	32.5	41.0	44.7	62.9	83.2	44.8	51.4	58.3	72.4	91.7	7.3	11.1	13.6	27.3	56.6
10	32.7	37.7	39.5	68.2	91.6	48.9	57.3	63.6	82.2	96.3	6.1	12.2	17.3	36.0	71.0
20	23.1	33.6	39.8	75.8	97.2	54.9	64.9	70.3	91.6	99.4	9.5	14.7	20.4	56.1	89.6
40	18.4	28.3	40.4	84.7	99.7	60.4	71.7	76.3	97.9	99.9	11.3	23.1	28.0	78.9	99.4
100	7.5	19.7	34.5	94.4	100	62.8	83.9	89.5	100	100	13.5	35.5	51.6	97.3	100
Population value = .2, 0															
6	46.2	66.5	75.7	97.5	99.9	61.4	78.6	85.3	98.9	100	16.2	32.6	42.7	85.9	99.7
10	45.9	76.8	83.2	99.1	100	67.1	86.4	91.4	99.8	100	16.8	42.9	55.0	96.4	100
20	46.9	80.1	90.9	100	100	73.8	93.9	97.4	100	100	21.5	61.3	78.4	99.9	100
40	46.8	90.1	97.3	100	100	84.9	98.8	99.6	100	100	33.0	84.1	94.7	100	100
100	45.0	98.5	99.9	100	100	93.5	100	100	100	100	53.0	99.7	99.9	100	100
Population value = .1, 0															
6	32.1	38.7	40.6	60.7	83.2	47.4	51.7	54.7	75.0	90.8	6.0	9.4	10.6	28.8	52.4
10	28.3	36.0	42.2	66.0	88.6	47.9	55.1	63.1	81.1	95.3	7.4	11.6	13.7	35.2	69.0
20	24.4	33.9	40.2	72.1	96.2	54.8	63.8	96.6	90.4	99.4	7.4	14.0	19.2	50.5	89.1
40	18.8	28.3	34.3	81.1	99.5	56.2	67.8	77.1	97.0	99.9	10.4	21.3	26.9	72.1	99.1
100	7.6	18.0	30.7	91.9	100	62.7	81.5	88.1	100	100	12.1	33.2	49.3	98.2	100

procedures met the criterion for concluding multiple populations was calculated across the replications. For the set of conditions where correlations were based on the same population mean, these results reflect Type I error. For those conditions where correlations were based on two population means, these results reflect power.

One final summary statistic bearing on the robustness of the S&H

method was the mean of the ratios of error to correlation variance, which is reported for the condition where samples were all based on the same population correlation. This mean reflects bias if it differs from 1.0 because under this condition (all correlations from one population) both variances should reflect only error. For convenience, we chose the population correlation value of zero to demonstrate the effect.

Results

The results of this Monte Carlo study are summarized in Tables 1–3. Table 1 concerns Type I error rates. It indicates the percentage of times the two S&H procedures and *U* statistic detected population differences when they did not in fact exist. As can be seen, the *U* statistic maintained an error rate of approximately .05 when chi-square was tested for significance at this level. Neither number of correlations compared nor sample size per correlation seemed to have an effect on the Type I error rate for the *U* statistic.

The results with the S&H–95 were reasonably consistent across conditions. Type I error rates were in the .35 to .45 range. The results with the S&H–75 procedure were inconsistent across conditions. Although the sample size per correlation and underlying population mean seemed to have a small effect, the number of correlations compared had a substantial effect on Type I error rates. When a relatively small number of correlations (6 or 10) were compared, the error rate was in excess of .20 most of the time. With 40 correlations, the error rate came closer to .10. With 100 correlations, the error rate dropped to about .02.

Table 2 presents the same information as Table 1, but for the 15 conditions where underlying population means were not the same. The ability to detect multiple populations when they do in fact exist reflects power, which is, of course, the converse of Type II error. As can be seen, both the S&H procedures and *U* were affected by the magnitude of difference between the population means, the sample size per correlation, and the number of correlations compared. For the most part, the bigger the differences, the more correlations compared, and the bigger the sample sizes, the more power. There were exceptions, however, for the S&H–75 procedure involving number of correlations compared when population differences differed by .1. Power at first increased and then decreased as the number of correlations compared increased (cf. Osburn et al., 1983). This tendency seemed somewhat obscured by a ceiling effect in conditions where power was close to 100%.

Table 3 summarizes the mean ratio of error to correlation variance for each of the 5 × 5 conditions when population correlations were all zero. Under these conditions, deviations from 1.0 reflect bias. As can be seen, with relatively few correlations compared, this ratio was quite far from 1.0. With relatively large numbers of correlations compared, the ratio approached 1.0.

Discussion

This Monte Carlo study was designed to ascertain the susceptibility of the Schmidt and Hunter meta-analysis procedure to Type I and Type II errors. The study was limited to situations where the only artifact in the observed correlations was sampling error. For comparison, the *U* statistic, a statistic for detecting statistically significant differences within a set of correlations, was also evaluated.

The Type I error rate was calculated by determining the percentage of time each method detected differences within a set of correlations that were in fact from the same population. The S&H–95 procedure resulted in unacceptable Type I error rates in all conditions; S&H–75 yielded less consistent results across

Table 3
Ratio of Error to Observed Variance When All Samples Are Based on Correlation Parameters of Zero

No. of <i>r</i> s	Sample size per <i>r</i>				
	30	75	100	250	500
6	1.58	1.67	1.67	1.78	1.66
10	1.20	1.24	1.26	1.27	1.31
20	1.06	1.10	1.10	1.11	1.11
40	1.00	1.04	1.04	1.04	1.05
100	0.99	1.00	1.01	1.01	1.02

different numbers of correlations compared. When less than 40 were compared, the error rate was unacceptably high, in contrast to the convention of 5%. Error rates as high as 32% were detected, and almost all were over 15%. The *U* statistic achieved the 5% error rate that it should, in theory, achieve. When the number of correlations reached 100, the S&H–75 procedure was less susceptible to Type I errors than was *U*.

The power of the S&H procedures was in most cases, greater than the *U* statistic, thus yielding a smaller Type II error rate. The increased power, however, was at the cost of inflated alpha error. As expected, S&H–95 showed greater power than S&H–75. Inspection of Table 2 indicates that *U* was actually more powerful than S&H–75 in many of the conditions involving comparisons of 100 correlations. These conditions were those where the S&H–75 procedure had the smaller Type I error rate.

Cohen (1977) has suggested a minimum acceptable level of power of .80, or a Type II error rate of .20. As can be seen in Table 2, many of the conditions had power well below this level. Unless the difference in population correlations is relatively large (e.g., .5 vs. .2), both S&H–75 and the *U* will probably not detect them unless sample sizes are considerable, certainly several hundred per study. The S&H–95 procedure had acceptable power under most conditions.

Overall, the Type I error rate for S&H–75 is unacceptably large under many conditions and for S&H–95 under all conditions. Furthermore, power of S&H–75 to detect differences is too small under many conditions. More troublesome, however, is that the means of ratios of error to correlation variance was inconsistent and too large. That the ratio decreased with increasing number of correlations compared, and that this decrease was accompanied by a decrease in the Type I error rate for S&H–75, both suggest a bias in the S&H procedure. The S&H formula appears not to be an unbiased estimate of population variance when the number of correlations is relatively small (see Raju & Burke, 1983). This may explain why adjustment of the correlation variance for artifacts often results in a negative variance. Of particular concern is that the corrections for other artifacts, in addition to sampling error, may increase the likelihood of Type II errors over what we have observed.

The results of this Monte Carlo study clearly show problems with the S&H procedure. It is not that the general approach is faulty, but rather that there are problems with some of the formulas to estimate variances. Unfortunately, doubt is now cast about conclusions in certain existing meta-analyses based on the procedure. Tables 1 and 2 can be used to help evaluate

conclusions in this prior research. Where the procedure has led to the conclusion of single populations, as in the validity generalization work of Schmidt and Hunter, the power tables can be utilized to indicate the likelihood that differences of a given size would have been detected. For studies suggesting the existence of moderators, such as Fisher and Gitelson (1983), the Type I error rate can be estimated. (These meta-analysts also used the *U* statistic and conclusions based on it may suffer from Type II errors, but probably do not suffer from Type I.)

Clearly, additional work needs to be done with the S&H procedure to work out unbiased estimates of variances, preferably through rigorous mathematical derivations, particularly when small numbers of correlations are compared. Caution is clearly indicated in circumstances where our analyses indicate unacceptably low power or high rates of Type I error, particularly because such errors are more serious for aggregations of studies than for single studies. In the interim, we recommend using our tables to estimate Type I and Type II error rates in deciding whether to use *U* or the S&H procedure. With large numbers of correlations compared, S&H-75 yields acceptable error rates and would represent a reasonable choice because it incorporates corrections for other artifacts. Where sampling error is the only artifact in question, or with small numbers, *U* would be preferable, although it does have power difficulties to detect small correlation differences.

References

- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.
- Fisher, C. D., & Gitelson, R. (1983). A meta-analysis of the correlates of role conflict and ambiguity. *Journal of Applied Psychology*, 68, 320-333.
- Glass, G. V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher*, 5, 3-8.
- Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, 96, 72-98.
- Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.
- Jackson, S. E. (1984, August). *Can meta-analysis be used for theory development in organizational psychology?* Paper presented at the meeting of the American Psychological Association, Toronto, Canada.
- Marascuilo, L. A. (1971). *Statistical methods for behavioral science research*. New York: McGraw-Hill.
- Osburn, H. G., Callender, J. C., Greener, J. M., & Ashworth, S. (1983). Statistical power of tests of the situational specificity hypothesis in validity generalization studies: A cautionary note. *Journal of Applied Psychology*, 68, 115-122.
- Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology*, 65, 373-406.
- Raju, N. S., & Burke, M. J. (1983). Two new procedures for studying validity generalization. *Journal of Applied Psychology*, 68, 382-395.
- Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage.
- SAS Institute. (1982). *SAS user's guide: Basics 1982 edition*. Cary, NC: SAS Institute.
- Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: A red herring. *Journal of Applied Psychology*, 66, 166-185.
- Spector, P. E. (1985). Higher-order need strength as a moderator of the job scope—employee outcome relationship: A meta-analysis. *Journal of Occupational Psychology*, 58, 119-127.

Received February 14, 1986
Revision received May 19, 1986 ■

A Decision-Theoretic Approach to the Use of Appropriateness Measurement for Detecting Invalid Test and Scale Scores

Fritz Drasgow and Elaine Guertler
University of Illinois

In psychological measurement it is important to determine when a particular examinee's test or scale score provides an invalid measure of the trait or attitude being assessed. In this article we present several quantitative indices that have been found to effectively identify some types of inappropriate scores. These measures, termed *appropriateness indices*, are all derived from item response theory. They are computed directly from the item responses that are combined to form the test or scale of interest; information from other scales or tests is not needed. A decision-theoretic approach to the use of appropriateness indices in selection decisions and theoretical research is introduced. An example is then presented to illustrate how researchers can use appropriateness indices. Finally, we discuss policy options that are available for dealing with individuals who are identified as having inappropriate scores.

The score of a particular individual on a psychological measurement instrument may be invalid even when the test has satisfactory measurement properties for the examinee population as a whole and for subpopulations within the overall population. Examinees' scores on achievement tests and aptitude tests are said to be spuriously high when, for example, they copy some answers from more talented neighbors or when they have been given the answers to some questions. A score on an attitude scale is spuriously high when a subject "fakes good." Spuriously low test and scale scores may be caused by alignment errors (answering, say, the 10th item in the space provided for 9th item, answering the 11th item in the space provided for the 10th item), language difficulties that hinder an examinee on mathematical word problems but not on items requiring only calculations, atypical educational programs that do not cover all of the topics that form standard curricula, poor motivation, or unusual interpretations of items. An *inappropriate score* on a test or scale refers to a score that is spuriously high or spuriously low.

How can individuals with inappropriate scores be identified? One approach to this problem requires specialized scales such as Ghiselli's (1960, 1963) "predictors of predictability" and "predictors of differential predictability." The scales assessing lying and faking in the Minnesota Multiphasic Personality Inventory (Hathaway & McKinley, 1951) provide other examples. People with high scores on such scales are flagged, to indicate that their scores may not be comparable to those of others. Decisions and predictions about these people are often made on the basis of other information, such as their scores on other tests or clinical observations.

Developing specialized scales to identify invalid scores is time consuming and expensive. Furthermore, research on predictors of predictability suggests that these specialized scales are often situation specific and cross validate poorly (e.g., Brown & Scott, 1966).

In this article we review a different approach to identifying inappropriate scores. This approach, called "appropriateness measurement" by Levine and Rubin (1979), uses only responses to items included on the test or scale and does not require the development of a specialized scale. From the item responses we compute an *appropriateness index* score, which is large in magnitude when an examinee's item responses provide contradictory evidence about the trait being measured. For example, the index will be large when an examinee correctly answers some hard items but responds incorrectly to some easy items. Thus, it is possible to identify low-ability examinees who have been given the answers to, say, 20% of the test items, and high-ability examinees who make alignment errors over blocks of items (Drasgow, Levine, & Williams, 1985).

Powerful methods for appropriateness measurement can be very useful to researchers and practitioners. For example, misclassification errors that result from inappropriate test scores can be expensive for test users and can have serious implications for test takers. People with spuriously high test scores may be selected for jobs or academic programs for which they are incompetent, and spuriously low test scores may cause deserving individuals to be denied professional and academic opportunities. One expensive alternative to appropriateness measurement is massive retesting: In New York, for example, 12,000 Nursing Board examinees were retested because officials suspected that widespread cheating had occurred ("July Licensing," 1979). Note that it would be necessary to use a secure test that has been properly validated when retesting.

Item Response Theory

The appropriateness indices described in this article are based on item response theory (IRT). Item response theory

We wish to thank James Drasgow, Charles L. Hulin, Michael V. Levine, Mary E. McLaughlin, Malcolm Ree, and Mary Roznowski for their help with this article.

Correspondence concerning this article should be sent to Fritz Drasgow, Department of Psychology, University of Illinois, 603 East Daniel Street, Champaign, Illinois 61820.

comprises a class of models that deal with the relation between an individual's responses to items on a test or questionnaire and the psychological trait measured by the instrument. In other words, IRT models the relation between an individual's standing on a latent characteristic (attitude, ability, etc.), usually denoted by the Greek letter theta (θ), and the item responses. The parameter θ and its estimate $\hat{\theta}$ are analogous to an individual's true and observed scores from classical test theory, respectively. In fact, θ is a monotonic (but nonlinear) function of true score when the usual IRT assumptions are correct.

Central to IRT is a mathematical function relating the probability of a particular response (e.g., a correct response to a multiple-choice item or a positively keyed item on an attitude scale) to an individual's standing on the latent trait. This relation can be graphically depicted in the form of an item-characteristic curve (ICC), which plots the probability of a correct response to the item for each level of θ . Frequently, an S-shaped curve such as the cumulative normal ogive provides a good description of these probabilities. A mathematically convenient S-shaped curve with a nonzero lower asymptote is the *three-parameter logistic ogive*. Here

$$P_i(\theta) = c_i + \frac{1 - c_i}{[1 + \exp\{-Da_i(\theta - b_i)\}]}, \quad (1)$$

where $P_i(\theta)$ is the probability of a correct response on the i th item among individuals with ability θ ; a_i , b_i , c_i are the item parameters; and D is a scaling constant usually set equal to 1.7. The *item difficulty* b_i corresponds to the point along the ability continuum at which examinees would have a 50% chance of a correct response if there were no guessing (i.e., if $c_i = 0$). The steepness of the ICC is indexed by the a_i parameter. It is called the *discriminating power* of the item. The *lower asymptote* c_i of the ICC is the probability of a positive response among individuals with very low θ values. The inclusion of this "guessing" parameter makes the three-parameter logistic model appropriate in situations where even very low-ability individuals occasionally respond correctly (e.g., on multiple-choice ability tests).

Item response theory parameters can be estimated using a computer program such as LOGIST (Wood, Wingersky, & Lord, 1976; Wingersky, Barton, & Lord, 1982) or BILOG (Mislevy & Bock, 1983). Initial work (Lord, 1968) on parameter estimation indicated that large samples (1,000 or more) and large numbers of items (50 or more) were required for IRT. More powerful estimation methods have since been developed (Bock & Aitkin, 1981; Rigdon & Tsutakawa, 1983) and shown to provide accurate parameter estimates with as few as 5 items and 200 examinees under some conditions (Drasgow, 1986).

Appropriateness Measurement

Item response theory provides a model-based approach to the identification of individuals for whom total test scores are not representative measures of their abilities. This approach examines whether an individual's pattern of item responses deviates from that expected on the basis of an IRT model.

Of course, we must first find an IRT model that provides a

reasonably good fit to normal response patterns before unusual response patterns can be identified. In previous research, the three-parameter logistic model has been found to be adequate for modeling the multiple-choice items on the Scholastic Aptitude Test-Verbal section (SAT-V; Drasgow et al., 1985; Levine & Drasgow, 1982) and the Graduate Record Examination-Verbal section (Drasgow, 1982; Levine & Drasgow, 1982). A special case of the three-parameter logistic model has $c_i = 0$ for all items. This model is called the *two-parameter logistic* model. It was found by Parsons (1983) to be satisfactory for modeling the Job Descriptive Index (JDI), an instrument developed by Smith, Kendall, and Hulin (1969) to assess five aspects of job satisfaction.

A number of appropriateness indices have been proposed in the literature. We describe several indices here that have been found to be good candidates for practical work (Drasgow, Levine, & McLaughlin, in press).

The first index is a standardized version of the logarithm of the likelihood of an examinee's responses evaluated at the maximum likelihood estimate $\hat{\theta}$ of the latent trait θ (Drasgow et al., 1985; Hulin, Drasgow, & Parsons, 1983, chapter 4). To compute this index,¹ denote a correct response to the i th item by $u_i = 1$, and an incorrect response by $u_i = 0$. The value $\hat{\theta}$ of θ that maximizes the likelihood of the observed item responses is first determined and then the likelihood is evaluated by computing

$$\ell_0 = \sum_{i=1}^n \{u_i \ln P_i(\hat{\theta}) + (1 - u_i) \ln [1 - P_i(\hat{\theta})]\}. \quad (2)$$

Standardization yields

$$\ell_z = \frac{\ell_0 - E(\ell_0)}{[\text{Var}(\ell_0)]^{1/2}}, \quad (3)$$

where

$$E(\ell_0) = \sum_{i=1}^n \{P_i(\hat{\theta}) \ln P_i(\hat{\theta}) + [1 - P_i(\hat{\theta})] \ln [1 - P_i(\hat{\theta})]\},$$

and

$$\text{Var}(\ell_0) = \sum_{i=1}^n P_i(\hat{\theta})[1 - P_i(\hat{\theta})]\{\ln[P_i(\hat{\theta})/(1 - P_i(\hat{\theta}))]\}^2.$$

The ℓ_z index is a measure of the goodness of fit (in the maximum likelihood sense) of an IRT model to a particular response pattern.

A second appropriateness index, described by Rudner (1983), is

$$F = \frac{\sum [u_i - P_i(\hat{\theta})]^2}{\sum P_i(\hat{\theta})[1 - P_i(\hat{\theta})]}. \quad (4)$$

Although some appropriateness indices are more complicated than F , they are all designed to identify inconsistencies in a response vector. With F it is particularly clear that inconsistent

¹ Fortran subroutines that compute the three indices described in this section can be obtained from the authors.

responses will cause the index to be large: The terms in the numerator of Equation 4 will be large when an examinee misses easy items ($u_i = 0$, and $P_i(\hat{\theta})$ is close to 1) and correctly answers items that are very difficult ($u_i = 1$, and $P_i(\hat{\theta})$ is close to 0).

A third index is the "fourth standardized extended caution index" proposed by Tatsuoaka (1984):

$$T4 = \frac{\sum [(P_i(\hat{\theta}) - u_i)(P_i(\hat{\theta}) - \bar{P})]}{[\sum P_i(\hat{\theta})(1 - P_i(\hat{\theta}))(P_i(\hat{\theta}) - \bar{P})^2]^{1/2}}, \quad (5)$$

where \bar{P} is the mean probability of correct responses at ability $\hat{\theta}$ over the n test items:

$$\bar{P} = \frac{1}{n} \sum_{i=1}^n P_i(\hat{\theta}).$$

This index provides a measure of the degree to which the pattern of item difficulties in the general examinee population covaries with the item difficulties experienced by a particular examinee as indicated by his or her pattern of right and wrong answers. $T4$ is an IRT appropriateness index that is similar in spirit to the "personal biserial" index developed by Donlon and Fischer (1968) to detect inappropriate test scores.

In practical applications one can modify each of the preceding equations by summing across only items answered by an examinee or subject. In particular, omitted and not-reached items do not have to be included in the summations. With multiple-choice test data, we have obtained improved rates of detection of simulated spuriously high and spuriously low response patterns with the ℓ_z index when omitted and not-reached items were excluded. It seems natural to modify F in this way also. With $T4$, however, summing over only answered items is more questionable (because \bar{P} would change), and consequently, D. Harnisch (personal communication, March 1984) suggested treating unanswered items as incorrects.

Information about aberrance can also be combined across several tests or scales (Drasgow et al., 1986). For example, ℓ_z has the multitest extension

$$\ell_z = \frac{\sum_{j=1}^m [\ell_0^{(j)} - E(\ell_0^{(j)})]}{[\sum_{j=1}^m \text{Var}(\ell_0^{(j)})]^{1/2}}, \quad (6)$$

where $\ell_0^{(j)}$, $E(\ell_0^{(j)})$, and $\text{Var}(\ell_0^{(j)})$ refer to the value of ℓ_0 , its expectation, and its variance, respectively, on the j th test or scale. It has been found that multitest versions of ℓ_z and $T4$ can provide substantial gains in the detection of aberrant response patterns.

Evaluating the Effectiveness of an Appropriateness Index

The classification of response vectors as aberrant or normal on the basis of appropriateness index values is never perfect; the distributions of index values associated with each type of response vector overlap to some extent, so that an observed index score may have been sampled from either distribution. Ap-

propriateness indices do, however, effectively classify response vectors, and research to date has focused primarily on assessing their accuracy.

Most of this research has been similar to the design devised by Levine and Rubin (1979) and extensively described by Hulin et al. (1983). Typically, the data (whether real or simulated) are divided into two samples, normal and aberrant. The normal sample consists of examinee response vectors that are thought to fit the IRT model under consideration. (It is only with simulated data that response vectors can be "truly normal"; if actual test data are used we can only assume this to be the case.) In experimental studies, aberrant response vectors are simulated by modifying the item responses of initially normal response patterns. A spuriously high response vector can be simulated by randomly selecting $k\%$ of the original responses and rescored them as correct. Spuriously low response vectors can be simulated by randomly selecting $k\%$ of the original responses and substituting random responses.

An appropriateness index is then computed for each response vector in both samples, and its effectiveness is evaluated by the degree to which it differentiates between normal and aberrant response vectors. Receiver operating characteristic (ROC) curves are often used to display graphically index effectiveness. Here the proportion of aberrant response vectors classified as aberrant (hits) is plotted against the proportion of normal response vectors incorrectly classified as aberrant (false alarms), for a number of values of the appropriateness index. This analysis is usually conducted separately for the spuriously high and spuriously low treatments because index accuracy may differ with regard to the two forms of aberrance (Drasgow, 1982; Drasgow et al., 1985). An ROC curve that rises sharply from the origin toward the upper left corner represents good detection of aberrance in that the index correctly identifies a large proportion of aberrant response vectors with correspondingly few false alarms. A detailed description of the use of ROC curves for evaluating appropriateness indices is given by Hulin et al. (1983).

Some types of aberrant response patterns are very detectable. For example, Drasgow et al. (in press) found that at a .1% false alarm rate (i.e., 1 false alarm per 1,000 normal examinees), the ℓ_z , F , and $T4$ indices detected 78%, 63%, and 69%, respectively, of simulated very low-ability SAT-V examinees ($-2.05 < \theta < -1.32$; θ s are usually scaled to have a mean of zero and a variance of one) subjected to the 30% spuriously high treatment. It is of course more difficult to detect cheating by examinees with higher abilities. The ℓ_z , F , and $T4$ indices detected only 31%, 31%, and 38%, respectively, of response patterns generated from slightly below average ($-0.52 < \theta < -0.05$) ability levels and subjected to the 30% spuriously high treatment.

A Decision-Theoretic Approach to the Use of Appropriateness Indices

An examination of the ROC curves obtained for different appropriateness indices is useful in determining whether they exceed minimum standards of index acceptability and in making tentative, graphic comparisons of the indices' accuracies. However, selecting an appropriateness index cutting score for some

Table 1
Two-by-Two Table Relating Decisions to Outcomes
and Their Associated Utilities

Decision	State of nature	
	Aberrant	Normal
Ab'	O_1 = hit U_1 = utility	O_2 = false positive U_2 = utility
N'	O_3 = miss U_3 = utility	O_4 = correct decision U_4 = utility

particular application requires a consideration not only of the degree to which the index can discriminate between normal and atypical response vectors, but also the relative value or importance to the user of the outcomes of the classification process. Note that the importance to an examinee with an aberrant test score may not coincide with importance to the user. This policy issue is considered later in this article.

Judgments of index effectiveness center on the trade-off between correctly identifying aberrant response vectors and avoiding the misclassification of normal response vectors, which in turn is a function of the utility or disutility assigned to these outcomes. In addition to these two considerations, a third factor, the prior probability (or base rate) of aberrance, must be taken into account when classifying response patterns on the basis of the appropriateness index selected for use.

The ensuing discussion can be simplified by depicting the problem confronted by the user in a 2×2 table formed by a combination of two factors: (a) the state of nature—whether the response vector is aberrant (Ab) or normal (N), and (b) the decision made by the user—classifying the response vector as aberrant (Ab') or normal (N'). The four outcomes of the classification process and their utilities are shown in Table 1.

The expected utility of concluding that a particular response vector is aberrant is

$$E_x(\text{Ab}') = P(\text{Ab}|x)U_1 + P(\text{N}|x)U_2, \quad (7)$$

and the expected utility of concluding that it is normal is

$$E_x(\text{N}') = P(\text{Ab}|x)U_3 + P(\text{N}|x)U_4, \quad (8)$$

where x is the value of the appropriateness index, and $P(\text{Ab}|x)$ and $P(\text{N}|x)$ are the probabilities that the response pattern is actually aberrant or normal, respectively, given the index value of x . Obviously, we should conclude that a response pattern is aberrant if $E_x(\text{Ab}') > E_x(\text{N}')$ and conclude that it is normal otherwise. The practical problem is to determine the values of $E_x(\text{Ab}')$ and $E_x(\text{N}')$ for each index score x so that one can determine which expected utility is larger.

As in other decision-making situations, we ordinarily have the reverse conditional probabilities, $P(x|\text{Ab})$ and $P(x|\text{N})$. These conditional probabilities can be obtained from simulation studies or experimental studies (instructing some examinees to fail deliberately, cheat from a neighbor, etc.). If $P(\text{Ab})$ is the base rate of aberrance in a particular situation, then a decision strategy that is mathematically equivalent to the one pre-

viously described is to conclude that a response pattern is aberrant if

$$\frac{P(x|\text{Ab})}{P(x|\text{N})} > \frac{1 - P(\text{Ab})}{P(\text{Ab})} \times \frac{U_4 - U_2}{U_1 - U_3}, \quad (9)$$

and normal otherwise. Equation 9 can be obtained from the relation $E_x(\text{Ab}') > E_x(\text{N}')$ by algebraic operations and an application of Bayes's theorem.

Equation 9 shows how base rates, utilities, and the likelihood ratio (the term on the left side of the inequality) interact in classification. Note that index scores that are more likely to have been computed from aberrant response vectors than from normal response vectors (i.e., $P(x|\text{Ab})/P(x|\text{N})$ is large) would usually be classified as aberrant. But the base rates of normals, $1 - P(\text{Ab})$, and aberrants, $P(\text{Ab})$, also enter into the classification decision, as do the utilities of each outcome.

It is tempting to select a single cutting score t for use with an appropriateness index and then classify as aberrant if $x > t$. Unfortunately, the use of a single cutting score for x can be inconsistent with Equation 9 when there are unequal base rates of normals and aberrants or unequal index variances in the two groups. Consequently, to classify index scores in a way that is consistent with Equation 9 it may be necessary to use several intervals. For example, when both large (say, +10 and larger) and small (say, -10 and smaller) index scores indicate aberrance we would classify as aberrant if x is in the interval $(-\infty, -10)$ or in the interval $(+10, +\infty)$. In both cases, $P(x|\text{N})$ would be small, $P(x|\text{Ab})$ would be relatively large, and their ratio $P(x|\text{Ab})/P(x|\text{N})$ would be large. Consequently, a single cutting score can be used with their ratio.

Estimation of Base Rates

In the preceding discussion we have assumed that the prior probability of aberrance (i.e., the base rate) is known. The estimation of the aberrant group's base rate, $P(\text{Ab})$, presents a peculiar problem in that the only way such an estimate can typically be made is through the use of the appropriateness index itself. But without knowledge of the prior probability of aberrance, Equation 9 cannot be used and the proportion of response vectors that are aberrant cannot be determined. Nonetheless, Rorer and Dawes (1982) provided a clever way of obtaining an estimate of $P(\text{Ab})$. They suggested the formula

$$\hat{P}(\text{Ab}) = \frac{P_t(\text{Ab}') - P_t(\text{Ab}'|\text{N})}{P_t(\text{Ab}'|\text{Ab}) - P_t(\text{Ab}'|\text{N})}, \quad (10)$$

where $\hat{P}_t(\text{Ab}')$ is the proportion of a representative sample of response vectors whose appropriateness index values exceed an arbitrary value t . Although the derivation of Equation 10 does not require the use of any particular value of t , it is probably a good idea to use a value that is reasonably close to a value that might be used in practice (e.g., -2.0 for ℓ_2). The other two probabilities in Equation 10, $P_t(\text{Ab}'|\text{Ab})$ and $P_t(\text{Ab}'|\text{N})$, denote the probabilities of hits and false positives when t is used as a cutting score. They can be obtained by means of a simulation study or by examining results presented by Drasgow et al. (in press).

Example

Tests and Data Set

To illustrate the use of appropriateness measurement, we describe some results obtained by Drasgow et al. (in press) and provide several additional analyses required to use appropriateness measurement in practical settings. The tests analyzed include the Arithmetic Reasoning (AR), Word Knowledge (WK), and Paragraph Comprehension (PC) tests from the Armed Services Vocational Aptitude Battery. A large sample of examinees who responded to these tests was collected in 1980; this data set is fully described in the *Profile of American Youth* (1982). The results presented here are based on two spaced samples of these examinees.

The American Youth sample consisted of high school students. Of course, the data from examinees who did not attempt to do their best can distort the results of any subsequent statistical analysis. This problem is analogous to the one encountered when uncooperative subjects respond halfheartedly to an attitude survey. Therefore, we wish to identify any American Youth examinees who provided spuriously low responses.

A large index score on any of the three appropriateness indices reviewed in this article points to aberrant responding, not just spuriously low responding: All three quantify inconsistency in a response pattern. We have found them to detect effectively simulated cheating by low-ability examinees and simulated spuriously low responding by high-ability examinees. Consequently, a large index score does not indicate the type of aberrance; it only reflects inconsistency in a response pattern (but see Drasgow et al., 1985, for an index that is much more effective for detecting spuriously low responses than spuriously high responses). American Youth examinees with large index scores shall be considered to have spuriously low scores because there is relatively little motivation to cheat on an experimental exam; the index score itself, however, does not justify this conclusion. If some examinees did cheat, we may do the right thing (discard their data) for the wrong reason (because we believed their responses were spuriously low).

Model Fit

Prior to any application of appropriateness measurement, we must first find an IRT model that provides a good fit to the observed item responses. We selected the three-parameter logistic model because the AR, WK, and PC tests are all multiple-choice tests that use number-right scoring. Hence, examinees should answer every item, which suggests that the lower asymptotes of ICCs will be nonzero. The fit of the three-parameter logistic model was assessed with the first spaced sample from the American Youth data set. This sample consisted of 2,978 examinees obtained by selecting the first, fifth, ninth, . . . response patterns.

One assumption of the three-parameter logistic model is unidimensionality. Here it is assumed that the latent trait θ is scalar valued (as opposed to being a vector). Although a perfectly unidimensional test of more than two items is probably impossible to obtain, it is important for a single dominant trait to underlie the items that form a given test or scale.

One approach to determining whether a test is sufficiently unidimensional for IRT is "modified parallel analysis" (Drasgow & Lissak, 1983). Here the eigenvalues of the reduced tetrachoric correlation matrix (largest off-diagonal correlations are used as communality estimates) are computed and compared to the eigenvalues of a simulation data set that is truly unidimensional. The eigenvalues of a number of simulation data sets were provided by Drasgow and Lissak. The IMSL Library 1 (1975) subroutine BECTR can be used to compute the tetrachoric correlations, and most statistical packages (e.g., SAS) have routines for computing eigenvalues.

The largest five eigenvalues of the 30-item AR test are 11.65, 1.36, 0.57, 0.47, and 0.35. This pattern of eigenvalues is similar to the pattern given in Panel B of Figure 4 by Drasgow and Lissak (1983), which indicates that the AR test is strongly unidimensional. A 50-item verbal test was formed by combining the 35 WK items with the 15 PC items. The first five eigenvalues of this test, denoted WKPC, are 22.99, 1.55, 0.81, 0.69, and 0.46. Again, this indicates that the test measures a single dominant trait.

The LOGIST computer program was used to estimate item and examinee parameters for the AR and WKPC tests. Default convergence criteria were used, and convergence was obtained within the default number of iterations.

After estimating item parameters it is important to check that the parametric form of the ICC is adequate. "Fit plots" can be used to this end. Here we divided the sample into 25 subsamples on the basis of estimated ability. The 4th, 8th, . . . , 96th percentiles of the standard normal were used as cutting scores to form the subsamples. Within each subsample, the proportion answering each item correctly was determined. The fit plot for an item consists of its estimated ICC, the proportion correct for each of the 25 subsamples, and a vertical line through each proportion indicating an approximate 95% confidence interval (i.e., ± 2 standard errors).

The fit plots for four AR items are shown in Figure 1. Panels a and b depict two of the three items that seemed to us to have the worst fits of all of the AR items (Item 2 was similar to Item 1 and is therefore not shown). An item with a more-or-less average fit is shown in Panel c, and one of the better fitting items is shown in Panel d. The results for the WKPC test were similar. In sum, with the exception of the lowest ability stratum for the first two AR items, the three-parameter logistic ICCs seemed to provide reasonably good fits to the observed proportions and, consequently, justify the use of this model for appropriateness measurement.

Estimation of Probabilities

Two simulation data sets were created to determine the probabilities needed in Equation 10. The first sample consisted of 4,000 normal response patterns generated using the AR and WKPC item parameter estimates obtained from the two LOGIST runs. Note that all of the response patterns in the "normal" sample are truly normal. The other sample consisted of 2,000 response patterns that were first generated as normal response patterns and then subjected to the 30% spuriously low treatment. This treatment was selected because its effects are

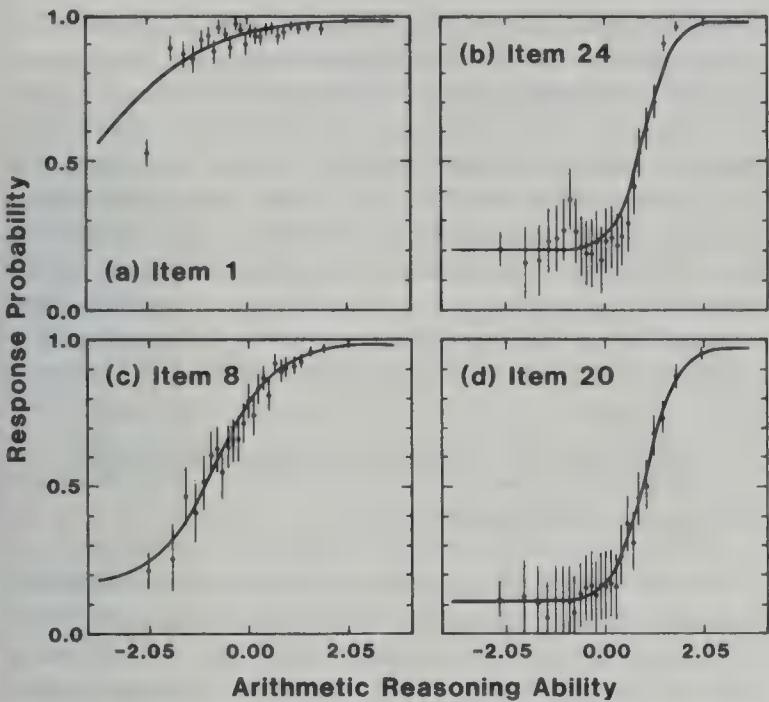


Figure 1. Empirical proportions, confidence intervals, and estimated item-characteristic curves for four Arithmetic Reasoning items.

large enough to alter significantly an examinee's score: About 15 responses would be changed for the 80 items on the two tests for an examinee of average ability.

The multitest ℓ_z index given in Equation 6 was computed for all response patterns. Its frequency distribution in the normal and aberrant samples is given in Figure 2.

The ROC curve for the simulated normals and simulated 30% spuriously low examinees is shown in Figure 3. Note that 42% of the aberrant sample can be identified when 1% of the normal response patterns are misclassified as aberrant and a 61% hit rate is obtained with a 5% misclassification rate. The

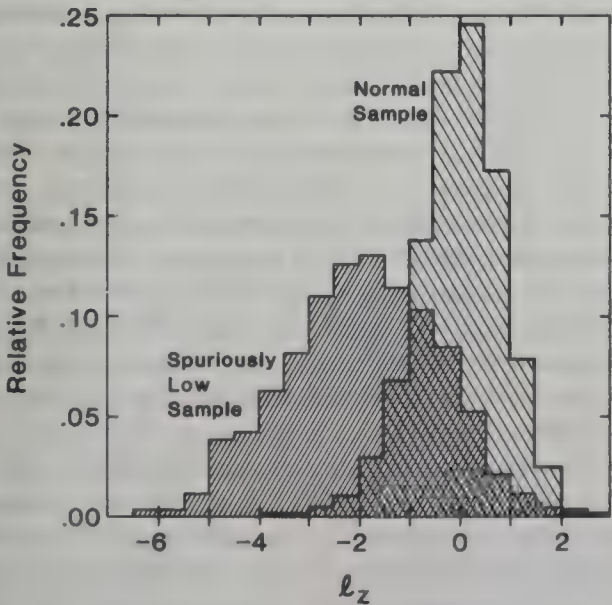


Figure 2. Relative frequencies of the ℓ_z appropriateness index in a sample of normal response vectors and in a 30% spuriously low sample.

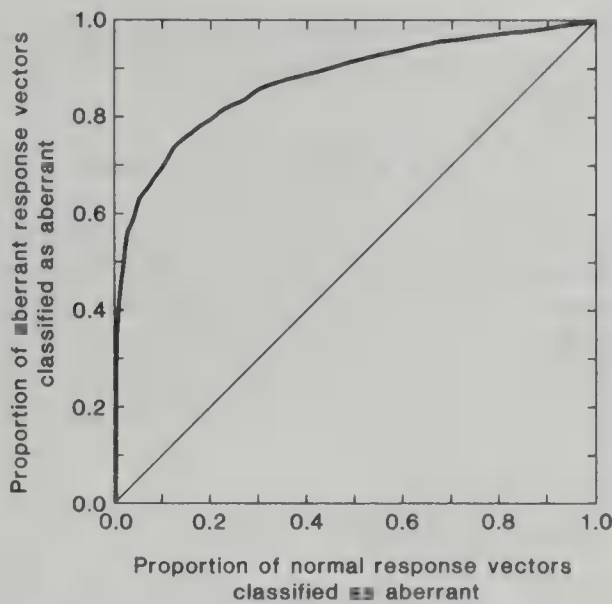


Figure 3. Receiver operating characteristic (ROC) curve for the multitest ℓ_z index computed for 4,000 normal response patterns and 2,000 response patterns subjected to the 30% spuriously low treatment.

value of ℓ_z used for the 1% error rate was -2.21 , and the ℓ_z value was -1.45 when the error rate was 5%.

At this point it is convenient to calculate two of the quantities required in Equation 10. From Figure 2 it is evident that relatively few normal response patterns have ℓ_z scores of less than -2.0 , so we decided to use this value to determine the probabilities in Equation 10. A total of 65 of the 4,000 normal response patterns had ℓ_z scores less than -2.0 , which yields the estimate

$$P_t(\text{Ab}'|\text{N}) = 65/4,000 = .016.$$

There were 960 ℓ_z scores of less than -2.0 in the aberrant sample, so that

$$P_t(\text{Ab}'|\text{Ab}) = 960/2,000 = .480.$$

Before calculating the optimal cutting score for ℓ_z we must estimate the base rate of aberrance in the American Youth sample and specify the four utilities in Table 1. To estimate the base rate, we selected a second spaced sample of 2,978 examinees (Examinees 2, 6, 10, . . .) from the original data base. Response patterns of examinees who provided answers to fewer than 77% of the items on either test were discarded because (a) such response patterns typically have very large appropriateness index scores (Drasgow et al., 1985), (b) on substantive grounds it seems clear that these response patterns are spuriously low (examinees should answer every item because there is no penalty for incorrect responses), and (c) a sophisticated IRT analysis is not required to identify these response patterns. The number of response patterns discarded due to excessive omitting was 262.

The ℓ_z index was then computed for the 2,716 examinees who answered at least 77% of the items on both tests. In this sample 93 examinees had index scores of less than -2.0 . Thus, we may estimate $P_t(\text{Ab}')$ from Equation 10 as $93/2,716 = .034$.

Inserting the needed quantities into Equation 10 yields

$$\hat{P}(\text{Ab}) = \frac{.034 - .016}{.480 - .016} = .039$$

as an estimate of the base rate of aberrance.

Of course, the estimate $\hat{P}(\text{Ab})$ should be used judiciously. From the analyses of fit previously described, we believe that the IRT model provides a good fit to the ASVAB data. Nonetheless, even minor violations of the assumptions of the IRT model might inflate $\hat{P}(\text{Ab})$ to some extent. Moreover, aberrant response patterns are sure to vary in the degree of their aberrance. Thus, $\hat{P}(\text{Ab})$ should be viewed as a heuristic statistic. Further research is needed in order to study optimal methods for estimating $\hat{P}(\text{Ab})$.

Utilities and Classification Rules

To classify response patterns as aberrant or normal, we must assign values to the utilities shown in Table 1. Although any set of utilities is arbitrary to some extent (see the discussion of policy issues in the next section), the following values can be used to illustrate the use of Equation 9.

First, it seems reasonable to set U_4 equal to zero inasmuch as no costs are incurred when a normal examinee is classified as normal. If the marginal cost of testing an additional examinee is \$75 (the examinees in the American Youth sample were paid an honorarium of \$50), then the utility U_2 might be assigned a value of -75 . We reach this conclusion because classifying a normal examinee as aberrant would lead us to discard the examinee's data and perhaps test an additional examinee. Notice that it is only the difference ($U_4 - U_2$) between these two utilities that is used in Equation 9; our selection of utilities is equivalent to a decision that it is worth \$75 to us to identify a normal response pattern as normal rather than as aberrant.

In the present example it is very difficult to assign a single value to the difference between U_1 and U_3 . Surely, the substantive conclusions of the American Youth study would be unaffected if a single aberrant response pattern were included in a very large sample of otherwise normal examinees. If more and more aberrant response patterns were included in the sample, however, substantive conclusions would be affected at some point. In this case the results of the study might mislead policy makers and cause serious planning errors. As a compromise between these two extremes, we set the difference $U_1 - U_3$ equal to 750, which is 10 times as large as $U_4 - U_2$.

Inserting the required quantities into Equation 9 shows that a response pattern should be classified as aberrant if

$$\frac{P(x|\text{Ab})}{P(x|\text{N})} > \frac{.961}{.039} \times \frac{75}{750} = 2.46.$$

From Figure 2 it is clear that $P(x|\text{Ab})$ is at least 2.46 times as large as $P(x|\text{N})$ when ℓ_z is less than -1.5 .

As a check on the sensitivity of this analysis to our choice of $U_1 - U_3$, we also evaluated Equation 9 using a difference of 300, which is four times the difference of U_4 and U_2 . Here we would be led to classify as aberrant when $P(x|\text{Ab})$ is at least 6.16 times as large as $P(x|\text{N})$. From Figure 2 it can be seen that we should classify as aberrant when ℓ_z is less than -2.0 . Thus, re-

ducing our evaluation of the cost of missing an aberrant response pattern from \$750 to \$300 causes us to be slightly more stringent in classifying a response pattern as aberrant.

Although no further analyses were conducted, one consequence of discarding data from high omitting examinees and low ℓ_z examinees is obvious. The cutting scores used to form the ability categories used by the military would certainly be changed. One important change is that the cutting score for the lowest ability group would in all likelihood increase; inasmuch as federal law prohibits enlisting examinees in this category, a smaller number of examinees would be eligible for the service.

Policy Issues in Appropriateness Measurement

Judgments of Outcome Utility

Selection and classification concerns can be expressed within the framework of the expected utility model by altering the relative importance accorded the outcomes in the payoff matrix shown in Table 1. Unfortunately, the problem of assigning utilities is not straightforward. For some examinees, inappropriate test scores can be used without decreasing the accuracy of decision making. For example, applicants with spuriously high test scores that nonetheless fall below the cutoff for selection will not be accepted, and applicants with spuriously low test scores that lie above the cutoff will still be accepted. No costs are incurred.

As is evident from the discussion in the last section, the process of assigning utilities to outcomes is complicated and likely to involve a number of subjective judgments. It is easy to disagree with the utilities that we have suggested. For example, the value of 750 assigned to the difference $U_1 - U_3$ might not be large enough in magnitude if the American Youth researchers were primarily concerned with nonrobust statistics such as means; it might be too large if robust statistics such as medians were of primary interest. Nonetheless, the example does indicate the broad outlines of the choices one must make to use appropriateness measurement in practice.

We expect that organizations concerned with personnel selection will ordinarily place a premium on detecting test scores that are spuriously high because hiring or admission decisions based on such scores may result in false positives (selected individuals who will not be successful). This may be particularly important in employment settings when testing costs are low but training costs are high. For example, the operational cost of administering the ASVAB is \$3.50 per person, whereas the cost of attrition from basic training across the Department of Defense is estimated to be \$12,500 per person (Malcolm Ree, personal communication, September 1985). Here it seems reasonable to assign a value to U_3 that is much larger in magnitude than any of the other utilities.

Organizations might place much less emphasis on detecting applicants who would be successful but might not be hired on the basis of their test scores. For example, in times of high unemployment an organization would not be likely to assign a value to U_3 that is much larger in magnitude than U_2 . (If there is no manpower shortage, then an organization will have little motivation to identify spuriously low scores.)

To handle the difference in importance accorded to the detection of spuriously high and spuriously low test scores, two payoff matrices would be required, each with its own utilities set to reflect costs and benefits. Ideally, two indices might be computed, one that detects spuriously high response patterns (but not spuriously low) and one that detects only spuriously low response patterns. Unfortunately, such indices are not yet ready for operational use, and so further research on this problem is needed.

In addition to the differing utilities assigned to the detection of spuriously high and spuriously low test scores, it is likely that the entries within each matrix will vary according to the situation (educational vs. employment), the use of the test score (research vs. applications, selection vs. placement or guidance), the cost of retesting, and whose utility function (the institution's, the applicants', or society's) is used. The latter distinction is likely to be of particular interest inasmuch as some (Hulin et al., 1983) view appropriateness measurement as a means for reducing the power differential between the testing organization and examinees by providing individual test takers with the means to dispute the validity of their test scores as measures of their abilities. This is viewed as being particularly important when decisions are based on these scores (i.e., hiring and college admissions) that may affect the individual's long-term welfare. It is, however, this very difference in power that makes it seem unlikely that applicants can impose their outcome preferences upon the institution. In all likelihood, utilities reflecting the value of detecting spuriously high, not low, scores will be used when selection is the primary purpose of testing. Of course, organizations concerned with long-term public policy ramifications of their actions could choose to detect both kinds of aberrance.

Actions to Be Taken Following Classification

The decision-making process does not end with the classification of a response vector as either aberrant or normal. Particularly when the test score in question is to be used to make subsequent decisions about the individual (e.g., whether to accept or reject an applicant), the decision maker must decide what action to take if the individual's test score has been found to be an inappropriate indicator of his or her ability.

What occurs when a test score is classified as inappropriate is likely to depend to a great extent on the user's inference as to the cause of the inappropriateness. In the case of a spuriously high test score, the cause of which may be cheating or some form of "coaching," in which the individual has access to part of the exam prior to its administration, the applicant may either be denied admission or be retested. Spuriously low test scores, on the other hand, frequently indicate the use of suboptimal test-taking techniques, which may not disappear if the applicant is simply retested. In these situations, some sort of active intervention may be necessary in order to teach the individual proper test-taking strategies (i.e., guessing when one or two distractors can be eliminated from consideration, not spending too long on any one item, etc.). Such interventions may be especially valuable in educational testing because it is probable that the individual will be taking many more tests in the future. Em-

ployers, on the other hand, may not find such a policy to be worth their while and resort to other means for making selection decisions (e.g., interviews and biographical information).

Conclusion

Appropriateness measurement accurately identifies some test-taking anomalies. In addition, appropriateness indices have been found to be effective when samples of several hundred examinees are available (Drasgow, 1982; much larger sample sizes are often thought to be required for IRT). Consequently, appropriateness measurement can be used in many applied situations.

In this article, we have outlined a general decision-theoretic approach to the use of appropriateness measurement. Even in cases where a formal decision theoretic analysis is not desired or possible, our discussion illustrates the importance of taking outcome utilities, base rates, and index accuracy into consideration.

An additional implication of the use of appropriateness measurement in industry and education is that policy must be developed for dealing with individuals for whom test scores are inappropriate. In all likelihood, these policies will be based, in part, on the user's hypotheses as to the causes of aberrance. This highlights the need to develop appropriateness indices sensitive to particular types of test-taking anomalies.

References

- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-459.
- Brown, F. G., & Scott, D. A. (1966). The unpredictability of predictability. *Journal of Educational Measurement*, 3, 297-301.
- Donlon, T. F., & Fischer, F. E. (1968). An index of an individual's agreement with group-determined item difficulties. *Educational and Psychological Measurement*, 28, 105-113.
- Drasgow, F. (1982). Choice of test model for appropriateness measurement. *Applied Psychological Measurement*, 6, 297-308.
- Drasgow, F. (1986). *Conditional and unconditional maximum likelihood estimation in item response theory*. Unpublished manuscript.
- Drasgow, F., Levine, M. V., & McLaughlin, M. E. (in press). *Appropriateness measurement* (AFHRL-TP-87). Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.
- Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38, 67-86.
- Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology*, 68, 363-373.
- Ghiselli, E. E. (1960). The prediction of predictability. *Educational and Psychological Measurement*, 20, 3-8.
- Ghiselli, E. E. (1963). Moderating effects and differential reliability and validity. *Journal of Applied Psychology*, 47, 81-86.
- Hathaway, S. R., & McKinley, J. C. (1951). *The Minnesota Multiphasic Personality Inventory*. New York: Psychological Corporation.
- Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Applications to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

- IMSL Library 1 (5th ed.). (1975). Houston, TX: International Mathematical and Statistical Libraries.
- July licensing exam invalidated in New York. (1979). *American Journal of Nursing*, 79, 1671, 1680, 1684.
- Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique, and validating studies. *British Journal of Mathematical and Statistical Psychology*, 35, 42-56.
- Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, 4, 269-289.
- Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 28, 989-1020.
- Mislevy, R. J., & Bock, R. D. (1983). Implementation of the EM algorithm in the estimation of item parameters: The BILOG computer program. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory/Computerized Adaptive Testing Conference* (pp. 189-202). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.
- Parsons, C. K. (1983). The identification of people for whom Job Descriptive Index scores are inappropriate. *Organizational Behavior and Human Performance*, 31, 365-393.
- Profile of American youth: 1980 nationwide administration of the Armed Services Vocational Aptitude Battery*. (1982). Washington, DC: Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics).
- Rigdon, S. E., & Tsutakawa, R. K. (1983). Parameter estimation in latent trait models. *Psychometrika*, 48, 567-574.
- Rorer, L. G., & Dawes, R. M. (1982). A base-rate bootstrap. *Journal of Consulting and Clinical Psychology*, 50, 419-425.
- Rudner, L. M. (1983). Individual assessment accuracy. *Journal of Educational Measurement*, 20, 207-219.
- Smith, P. C., Kendall, L., & Hulin, C. L. (1969). *The measurement of satisfaction in work and retirement*. Chicago: Rand McNally.
- Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika*, 49, 95-110.
- Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide*. Princeton, NJ: Educational Testing Service.
- Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). *LOGIST—A computer program for estimating examinee ability and item characteristic curve parameters* (Research Memorandum No. 76-6). Princeton, NJ: Educational Testing Service.

Received February 18, 1986

Revision received June 25, 1986 ■

Study of the Measurement Bias of Two Standardized Psychological Tests

Fritz Drasgow
University of Illinois

Psychological tests are subject to two distinct forms of bias. The first form, measurement bias, occurs when individuals with equal standing on the trait measured by the test, but sampled from different subpopulations, have different expected test scores. Relational bias, the second type of bias, exists with respect to a second variable if a measure of bivariate association differs across groups. Empirical studies have found little evidence of relational bias. Two recent court cases, however, seem to have been more influenced by considerations of measurement bias than by the literature concerning relational bias. Unfortunately, a consequence of both court cases is that the respective test makers must select items for future tests on the basis of a statistic (proportion correct) that is inappropriate for evaluating measurement bias. More sophisticated approaches may also suffer from methodological difficulties unless special precautions are taken. In this article, tests of English and Mathematics Usage are analyzed by measurement bias methods in which several steps are taken to reduce methodological artifacts. Many items are found to be biased. Nonetheless, the sizes of these effects are very small and no cumulative bias across items is found.

It is important to study the extent to which two distinct properties of a psychological test are satisfied when it is used in a heterogeneous population. The first property, *measurement equivalence*, holds when individuals with equal standing on the trait measured by the test but sampled from different subpopulations have equal expected observed test scores. *Biased measurement* occurs when a test fails to provide measurement equivalence. The second property of a test, *relational equivalence*, holds with respect to another variable if a bivariate measure of their association is identical across relevant subpopulations.

Studies of relational equivalence in employment and educational testing commonly use measures of performance on the job or in the classroom as the second variable. The bivariate relation that should be studied in these cases is the regression of performance on test score, including intercepts, slopes, and standard errors of estimate. We assume throughout this article that regression analyses using measures of performance are conducted to study relational equivalence. This general approach to relational equivalence is called *differential prediction*. (See Petersen & Novick, 1976, for a review and critique of alternative methods for assessing fairness in test use.)

Although a detailed review of the differential prediction literature is beyond the scope of this article, the general findings can be quickly summarized. Linn (1982) stated,

Whether the criterion to be predicted is freshman GPA in college, first year grades in law school, outcomes of job training, or job performance measures, carefully chosen ability tests have *not* been found to underpredict the actual performance of minority group persons. Contrary to what is often presupposed, the bulk of the evidence shows either that there are essentially no differences in predictions based on minority or majority group data, or that predictions based on majority group data give some advantage to minority group members. (p. 384)

Studies of sex differences have found that regression equations derived from men's data underpredicted women's performances as undergraduates (Linn, 1973) and in clerical military training courses (Dunbar & Novick, 1985), but not as first-year law school students (Pitcher, 1974, 1975).

The "testing controversy" has continued to play a significant role in American society despite the relatively coherent set of findings obtained by differential prediction researchers. Two recently settled legal cases are particularly noteworthy. In the first case, the Golden Rule Insurance Company filed suit against the Educational Testing Service (ETS) and the Illinois Department of Insurance (*Golden Rule Insurance Company et al. v. Washburn et al.*, 1984) when it was found that blacks had higher failure rates than whites on various Illinois insurance licensing exams constructed by ETS. In an out-of-court settlement, ETS made no admission of guilt with respect to any allegation by Golden Rule. Nonetheless, to end the litigation ETS agreed to several modifications in its procedures for constructing the licensing exams. One of these changes involves the way in which items are selected for the exams. The Educational Testing Service agreed to select items for which the proportions of correct answers for whites and blacks differ by no more than .15 when such items are available. In the second legal case, *Allen v. Alabama State Board of Education* (1985), the defendant agreed to even more stringent conditions concerning item selection. Here, items with proportions of correct answers for whites and blacks that differ by no more than .05 are to "be used exclusively as

I wish to thank Mark Reckase for providing the data analyzed here, Mary Roznowski for help with the data analysis and editorial comments, David Harrison for providing a simulated data set, and Charles Hulin and Lloyd Humphreys for their comments on earlier versions of this article.

Correspondence concerning this article should be sent to Fritz Drasgow, Department of Psychology, University of Illinois, 603 East Daniel Street, Champaign, Illinois 61820.

scoreable items so long as they are available in sufficient numbers to provide comprehensive coverage of the objectives sought to be measured in the examination" (*Allen v. Alabama State Board of Education*, 1985, consent decree, p. 4). It is important to note that there is no scientific justification for the test construction procedures described; they can only be justified on political grounds.

It is difficult to predict the ultimate impact of these cases on the construction and use of other standardized tests. The 1985 chair of the American Psychological Association Committee on Cultural Diversity in Testing was cited as having the opinion that the Golden Rule agreement "will generate considerable pressure on ETS, including legal pressure to use similar procedures for choosing items on other tests" and that "some are already thinking about requesting testing companies to operate in line with the [Golden Rule] agreement" (Cordes, 1985, p. 26). Subsequently, the New York Legislature has begun to consider extending its influential testing law so that the Golden Rule procedure for test construction would apply to standardized professional and occupational tests (Jaschik, 1986).

The plaintiffs in the two legal cases seemed to be very concerned with the issue of measurement bias. In part, this was due to the lack of data on criterion performance, which excluded the possibility of criterion-related validity studies. But in both cases the events that triggered the litigation were the discrepancies in passing rates for blacks and whites. These discrepancies led to the conclusion that "the tests are culturally biased" (*Allen v. Alabama State Board of Education*, 1985, consent decree, p. 2); the remedies in both cases involved selecting items with nearly equal proportions correct for whites and blacks. Unfortunately, comparing proportions of correct answers across subpopulations is a scientifically inadequate method for evaluating measurement bias (Hulin, Drasgow, & Parsons, 1983; Hunter, 1975) because the proportion-correct statistic confounds measurement bias with between-group differences in the attribute measured by the test.

Lord's (1977, 1980) method for studying measurement bias has a stronger statistical justification. (For reviews of methods for studying measurement bias, see Hulin et al., 1983, chapter 5; Petersen, 1980; and Scheuneman, 1980.) Using his method, which is based on item response theory (IRT), Lord (1977) found that 38 of the 85 items on the Scholastic Aptitude Test, Verbal section, were biased, when he analyzed the responses of 2,250 whites and 2,250 blacks. A similar result was obtained by Scheuneman (1982); she found that up to one half of the items written in some formats were biased, when she analyzed black and white subjects' responses to the Metropolitan Readiness Tests and the Otis-Lennon School Abilities Test.

Several methodological difficulties can occur in IRT studies of measurement equivalence. For example, it is usually necessary to make strong assumptions about the dimensionality of an item pool and the parametric form of the item-characteristic curves. If these assumptions are not reasonably satisfied, then violations of assumptions may cause truly unbiased items to appear biased in statistical tests. The linking of ability metrics, which is required prior to comparisons of item-characteristic curves, is also based to some extent on heuristic principles rather than on statistical theory. It is unknown how this affects the sampling distribution of bias statistics.

A subtle but serious problem with some IRT item bias studies has recently been identified by McLaughlin (1986). In a simulation study, McLaughlin generated 100 replications of 1,000 examinees responding to a 50-item test. Then the LOGIST computer program (Wood, Wingersky, & Lord, 1976) was used to estimate item and person parameters simultaneously for each of the 100 replications. Finally, 50 IRT item bias analyses were conducted: Replication 1 was compared to Replication 2, Replication 3 was compared to Replication 4, and so forth. The average proportion of items found to be biased by Lord's (1980, chapter 14) chi-square statistic was .12 when $\alpha = .05$, and the rejection rate was as high as 30% for some individual items. (This is more serious than the usual problem with multiple significance tests. Instead of the 2.5 Type I errors that are expected per 50 tests when $\alpha = .05$, McLaughlin found an average of 6 Type I errors per 50 tests and as many as 15 for some items.) McLaughlin's actual rejection rate was 4 times the nominal rate when $\alpha = .01$ and almost 10 times the nominal rate when $\alpha = .001$. When McLaughlin held the ability parameters fixed at their true values (which is impossible to do in practice) so that the assumptions of Lord's chi-square statistic were (for the most part) satisfied, the average observed rejection rate was .039, which is much closer to the nominal alpha level of .05. In sum, McLaughlin's research shows that we should expect to find many more items to be "biased" as a result of Type I errors than would be expected on the basis of the nominal alpha level of the test.

Finally, in employment and educational testing it should be emphasized that a measurement bias analysis is only the first of two analyses that are needed to examine potential bias (Drasgow, 1984; Petersen, 1980). The second analysis should examine possible differential prediction of the criterion variable. Due to the growing differential prediction literature and advances in meta-analysis, it may be unnecessary (and even misleading if only small samples are available) to actually conduct a differential prediction study within a particular organization. Nonetheless, it is essential to marshal evidence about differential prediction in addition to the analysis of measurement bias.

Purpose of the Study

Although there is little empirical evidence of predictive bias, there is abundant (and perhaps misleading) evidence of measurement bias. The Golden Rule and *Allen v. Alabama* cases seem to have hinged on issues concerning measurement bias instead of the literature on predictive bias. Unfortunately, both cases used an inappropriate method (proportion correct) to assess measurement bias. There will be serious consequences for employment testing if this inappropriate approach to measurement bias becomes mandated by the courts and legislatures. Therefore, the purpose of this study is to examine the measurement bias of two standardized tests using methods that are psychometrically justified.

The methods used in this article include three features designed to minimize methodological artifacts. First, a very stringent alpha level was selected to test for item bias. The nominal level selected was .0005, which McLaughlin (1986) found to yield an actual rejection rate of .0056 in three-parameter logistic analyses of 1,000 simulated subjects responding to 50-item

tests. This latter rejection rate should keep the total number of Type I errors across all significance tests at a reasonable level. The power of our tests should be high despite the stringent alpha level because all samples are quite large (about 1,500 per sample).

Iterative linking of ability metrics is a second aspect of the analyses designed to minimize methodological artifacts. Here item parameter estimates are initially placed on approximately equal scales by using all items, and then item bias statistics are computed. Subsequently, item parameter estimates are re-linked using only those items found to be unbiased, and again item bias statistics are computed for all of the items. This process continues until the same set of items is found to be biased on two successive iterations. The virtue of the iterative procedure is that items initially found to be biased are not used when metrics are re-linked. If biased items are not discarded, linking methods may compensate for truly biased items by causing truly unbiased items to appear biased (Segall, 1982). Iterative linking is similar in spirit to the test purification process suggested by Gary Marco and described by Lord (1980, pp. 220–221).

The iterative linking of ability metrics allows us to improve the analyses in a third way, namely by calculating expected number-right scores as a function of ability for each subpopulation. These values are very important because they show the net effect of item bias. Suppose, for example, that on a 50-item test a minority group member had an expected number-right score that was five less than a majority group member of the same ability. In such a case it would be clear that the test provided biased measurement and use of the test would be a source of serious concern to minority group members. Thus, calculation of expected observed scores as a function of the latent trait allows us to go beyond a piecemeal, item-by-item analysis and answer the fundamental question of whether the test as a whole provides biased measurement.

Method

Subjects and Measures

The data used in the analyses were a subset of the item responses of 140,979 examinees to the October 29, 1983 national administration of the American College Testing (ACT) Assessment. The ACT Assessment includes tests of English Usage, Mathematics Usage, Social Studies

Reading, and Natural Sciences Reading. Other data available were examinees' responses to the Student Profile section, an instrument with 190 questions concerning educational plans, demographic information, and personal interests.

Sex and racial information from the Student Profile section were used to form six data sets. Stratified samples of 1,500 white men and 1,500 white women were formed by selecting every 10th white man and every 10th white woman. Samples of 1,500 black men, black women, and Hispanic women were formed by selecting the first 1,500 examinees on the magnetic tape with these characteristics. Every Hispanic man on the tape was selected to form the 6th data set; there were 1,441 such examinees available.

Files with the dichotomously scored (1 if correct, 0 otherwise) item responses from the Mathematics Usage and English Usage tests were then created for each group. There were 12 files containing dichotomously scored item responses (2 tests, 6 sex-by-race groups). Finally, to conduct the item bias analysis as described by Lord (1980), files containing the item responses of 3,000 examinees to the Mathematics Usage and English Usage tests were created. A total of 500 examinees from each group except the Hispanic male group were selected by taking every third examinee. Sampling every third Hispanic man yielded only 480 ($= 1,441/3$) examinees; consequently, Examinees 2, 5, . . . , 59 from the Hispanic male file were also selected.

Analyses

Calibration and fit. The LOGIST computer program was used to estimate item and person parameters of the three-parameter logistic model for the two data sets with $n = 3,000$, and 12 data sets with $n = 1,500$, examinees. Default parameters were used to run the program. The lower asymptote parameters (c_i) were held fixed in the runs with $n = 1,500$ at the values obtained in the corresponding $n = 3,000$ runs (see Lord, 1980, p. 217). The dimensionality assumption of the three-parameter logistic model was checked by computing eigenvalues of the reduced matrix of tetrachoric correlations for each of the two tests using the samples of $n = 3,000$ and $n = 1,500$ and comparing them to simulation results. Largest off-diagonal correlations were used as communality estimates. The parametric form of the three-parameter logistic model item-characteristic curve was evaluated graphically by plotting estimated item-characteristic curves and observed proportions of examinees correctly answering the item in 25 ability strata. The ability strata were formed by first computing maximum likelihood ability estimates and then using the 4th, 8th, . . . , 96th percentile points of the standard normal distribution as cutting scores (see Hulin et al., 1983, pp. 19–22). The normal approximation to the binomial was then used to construct approximate 95% confidence intervals (± 2 estimated standard errors) for each of the observed proportions.

Table 1
Eigenvalues of Reduced Tetrachoric Correlation Matrices for the ACT Mathematics Usage Test

Sample	<i>n</i>	Eigenvalue no.						
		1	2	3	4	5	6	7
Mixed (men and women)	3,000	12.32	1.25	0.88	0.51	0.42	0.37	0.33
White men	1,500	13.37	1.31	0.88	0.56	0.47	0.43	0.40
White women	1,500	11.49	1.42	1.07	0.60	0.56	0.50	0.42
Black men	1,500	9.58	1.51	0.87	0.62	0.52	0.45	0.44
Black women	1,500	9.24	1.53	0.96	0.62	0.55	0.49	0.45
Hispanic men	1,441	12.26	1.44	1.00	0.59	0.55	0.50	0.42
Hispanic women	1,500	10.77	1.24	0.87	0.58	0.52	0.46	0.43

Note. Largest off-diagonal correlations were used as communality estimates for the 40 items. ACT = American College Testing Assessment.

Table 2
Proportions of Correct Responses to Mathematics Usage Items for Race and Sex Groupings

Group							Group						
Item	WM	WF	BM	BF	HM	HF	Item	WM	WF	BM	BF	HM	HF
1	.87	.83	.69	.66	.78	.74	21	.66	.54	.38	.29	.49	.38
2	.71	.62	.43	.41	.60	.51	22	.69	.59	.44	.44	.60	.49
3	.74	.69	.51	.51	.66	.59	23	.44	.35	.29	.26	.41	.30
4	.72	.64	.59	.53	.65	.59	24	.58	.40	.31	.21	.45	.32
5	.70	.57	.47	.41	.62	.51	25	.54	.44	.31	.32	.43	.38
6	.74	.61	.48	.42	.62	.48	26	.47	.44	.33	.35	.40	.40
7	.74	.64	.47	.40	.63	.52	27	.63	.52	.33	.29	.48	.37
8	.79	.70	.52	.49	.69	.56	28	.45	.35	.25	.24	.36	.29
9	.73	.67	.54	.51	.66	.58	29	.57	.43	.31	.28	.45	.35
10	.77	.62	.43	.32	.63	.50	30	.45	.37	.24	.23	.33	.29
11	.60	.63	.35	.42	.53	.52	31	.44	.38	.28	.25	.38	.31
12	.81	.72	.50	.49	.68	.60	32	.50	.42	.29	.28	.46	.32
13	.55	.49	.37	.35	.50	.40	33	.42	.33	.23	.23	.33	.27
14	.75	.60	.47	.39	.63	.47	34	.39	.31	.23	.24	.30	.30
15	.73	.60	.49	.43	.63	.48	35	.41	.29	.18	.18	.31	.21
16	.65	.58	.42	.39	.54	.49	36	.46	.27	.17	.12	.28	.16
17	.72	.55	.46	.41	.63	.50	37	.53	.40	.30	.28	.38	.33
18	.67	.55	.38	.34	.56	.46	38	.40	.35	.25	.25	.36	.31
19	.56	.53	.40	.36	.47	.43	39	.35	.28	.18	.18	.25	.21
20	.63	.47	.34	.25	.53	.34	40	.31	.24	.18	.17	.23	.20
No. of items failing 15% rule:							— 4 36 36 2 25						

Note. WM = white male; WF = white female; BM = black male; BF = black female; HM = Hispanic male; HF = Hispanic female.

Linking ability metrics and item bias statistics. Iterative Stocking and Lord (1983) linking was used to place estimated item parameters on the white male metric. In this procedure, Stocking and Lord linking was first used to link metrics. (The 99 percentile points of the standard normal distribution were used in place of estimated abilities to reduce computer expenses.) Then Lord's chi-square item bias statistic was computed for each item. Items found to be biased ($\alpha = .0005$) were

temporarily set aside, and only items found to be unbiased were used to relink the ability metrics. Then item bias statistics were recomputed for all of the items (including items found to be biased in the previous iteration). This process continued until the same set of items was found to be biased on two successive iterations.

Test-characteristic curves. To evaluate the cumulative effects of biased items, test-characteristic curves were computed for each group and each test. These curves, which are the sum of the item-characteristic curves, give the expected observed number-right score as a function of the trait measured by the test.

Results

Mathematics Usage Test

Table 1 presents the eigenvalues of the reduced tetrachoric correlation matrices for the Mathematics Usage test. The patterns of the eigenvalues for black men and black women seem to be quite similar to Figure 4b from Drasgow and Lissak (1983, p. 370). This indicates that the data sets are probably multidimensional but that most items are strongly related to a single, dominant general factor. Consequently, the parameter estimates computed by LOGIST will be almost as accurate as estimates obtained from a truly unidimensional data set (Drasgow & Parsons, 1983). The eigenvalues of the other groups also indicate that item parameters can be accurately estimated. These comparisons therefore justify our use of a unidimensional latent trait model.

Figure 1 illustrates the parametric fits of several Mathematics Usage items for the sample of $n = 3,000$. It shows estimated item-characteristic curves, empirical proportions of correct responses in various ability strata, and approximate 95% confi-

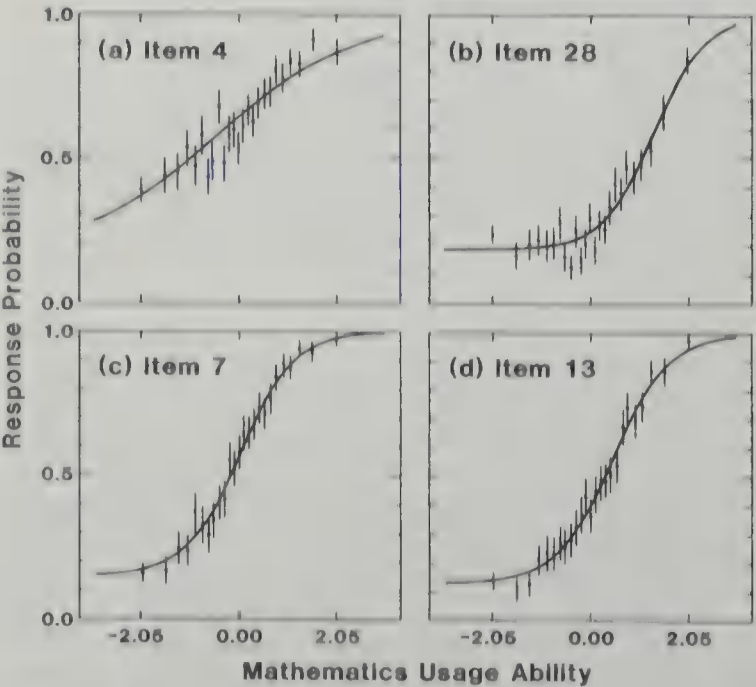


Figure 1. Empirical proportions, confidence intervals, and estimated item-characteristic curves for four Mathematics Usage items.

Table 3
Chi-Square Statistics for Mathematics Usage Items Found to Be Biased in Comparisons With White Men

Group							Group						
Item	WF	BM	BF	HM	HF	Total	Item	WF	BM	BF	HM	HF	Total
1		29 ^e	26 ^e		26 ^e	3	21			32 ^h	16 ^h	21 ^h	3
2						0	22		38 ^e	61 ^e		24 ^e	3
3						0	23						0
4						0	24	21 ^h		25 ^h			2
5			21 ^h			1	25						0
6						0	26	24 ^e	40 ^e	77 ^e		63 ^e	4
7			20 ^h			1	27		21 ^e				1
8						0	28						0
9		65 ^e	43 ^e		37 ^e	3	29						0
10		16 ^h	75 ^h			2	30	20 ^e	16 ^h			44 ^e	3
11	27 ^e					1	31					25 ^e	1
12						0	32			22 ^e		34 ^e	2
13		45 ^e	79 ^e	17 ^e	43 ^e	4	33		19 ^e				1
14	19 ^h		39 ^h		33 ^h	3	34						0
15						0	35						0
16	17 ^e	24 ^e	22 ^e		45 ^e	4	36	25 ^h		52 ^h	18 ^h	42 ^h	4
17	19 ^h					1	37						0
18						0	38	21 ^e		27 ^e	18 ^e	58 ^e	4
19						0	39					21 ^e	1
20			44 ^h		19 ^h	2	40						0
No. significant within group:							—	9	10	16	4	15	—

Note. WF = white female; BM = black male; BF = black female; HM = Hispanic male; HF = Hispanic female. ^e = easier (\hat{b}_i for indicated group is less than \hat{b}_i for white men); ^h = harder (\hat{b}_i for indicated group is larger than \hat{b}_i for white men). Insignificant statistics omitted.

dence intervals for the proportions. Figure 1a and Figure 1b, which show the results for Items 4 and 28, respectively, were selected because in our judgment they were among the items most poorly fit by three-parameter logistic item-characteristic curves. Item 4 is badly fitted because four of the 95% confidence intervals lie entirely below the estimated item-characteristic curve for ability strata near zero, and four intervals lie entirely

above the curve for average to high ability strata. Item 28 is similar in that four confidence intervals lie above the curve and three lie below. The lack of fit for these two items does not seem to be systematic: No remedy for the poor fit is obvious.

Items 7 and 13, presented in Figure 1c and Figure 1d, show fits that are typical of the vast majority of Mathematics Usage items. Only a few confidence intervals do not intersect the item-characteristic curve estimated for Items 7 and 13; the errant confidence intervals are nonetheless close to the estimated curves. We stress that these two items were not selected because they had the best fits. In fact, other items had all their confidence intervals crossing the estimated item-characteristic curve. Items 7 and 13 were selected for Figure 1 because their fits seemed to be average among the 40 Mathematics Usage items. Because the average fit of three-parameter logistic item-characteristic curves seemed to be reasonably good, we proceeded to the measurement equivalence analyses.

Proportions passing each of the Mathematics Usage items for the six sex-by-race groups are shown in Table 2. Note that 90% of the items fail to satisfy the 15% criterion enunciated in the Golden Rule case when the black male and black female proportions are compared to the white male proportions. Thus, ■ massive revision of this test would be necessary according to the Golden Rule criterion.

Is the Mathematics Usage test truly biased? Because it confounds bias with between-group differences, results of analyses based on the proportion-correct statistic are not trustworthy. The IRT analysis in contrast does not confound bias and between-group differences and consequently allows us to examine the measurement bias of the Mathematics Usage test in a rigorous way.

Table 4
Expected Number-Right Score as a Function of Ability for the Mathematics Usage Test

Ability	Group					
	WM	WF	BM	BF	HM	HF
-2.5	7.5	7.7	7.9	7.5	7.6	7.9
-2.0	8.9	9.0	9.4	8.8	9.0	9.2
-1.5	11.1	11.2	11.8	10.9	11.2	11.4
-1.0	14.4	14.5	15.3	14.1	14.6	14.8
-0.5	18.9	18.8	19.8	18.5	18.8	19.2
0.0	23.9	23.8	24.7	23.7	23.6	24.5
0.5	29.0	29.0	29.5	28.9	28.9	30.0
1.0	33.4	33.5	33.6	33.2	33.7	34.3
1.5	36.4	36.5	36.4	36.1	36.8	36.9
2.0	38.1	38.0	38.0	37.8	38.3	38.2
2.5	39.0	38.9	38.8	38.7	39.0	39.0

Note. All ability metrics were transformed to the white male metric, which was scaled to have a mean of zero and a standard deviation of one for the white male sample. WM = white male; WF = white female; BM = black male; BF = black female; HM = Hispanic male; HF = Hispanic female.

Table 5
Estimated Item Difficulties and Standard Errors for Items Found to Be Biased in at Least Three Comparisons

Item	Group					
	WM	WF	BM	BF	HM	HF
1						
\hat{b}	-1.49	-1.78	-2.05	-1.83	-1.56	-2.09
SE	0.08	0.08	0.07	0.05	0.07	0.09
9						
\hat{b}	-0.68	-0.89	-1.16	-1.09	-0.85	-1.07
SE	0.05	0.05	0.04	0.04	0.05	0.05
13						
\hat{b}	-0.07	-0.15	-0.39	-0.46	-0.20	-0.31
SE	0.04	0.05	0.06	0.04	0.05	0.04
14						
\hat{b}	-1.12	-0.62	-0.77	-0.28	-0.84	-0.43
SE	0.10	0.06	0.07	0.09	0.07	0.07
16						
\hat{b}	-0.32	-0.54	-0.61	-0.61	-0.37	-0.68
SE	0.04	0.04	0.04	0.04	0.04	0.04
21						
\hat{b}	-0.44	-0.31	-0.23	0.00	-0.09	-0.06
SE	0.07	0.05	0.08	0.07	0.06	0.06
22						
\hat{b}	-0.51	-0.63	-0.88	-0.96	-0.70	-0.80
SE	0.04	0.04	0.04	0.04	0.04	0.04
26						
\hat{b}	0.44	0.23	0.12	-0.16	0.36	-0.05
SE	0.04	0.06	0.10	0.08	0.07	0.07
30						
\hat{b}	0.46	0.24	0.48	0.37	0.43	0.07
SE	0.04	0.04	0.11	0.10	0.05	0.05
36						
\hat{b}	0.32	0.68	0.61	0.99	0.57	0.80
SE	0.04	0.07	0.10	0.14	0.05	0.09
38						
\hat{b}	0.85	0.59	0.90	0.40	0.61	0.32
SE	0.05	0.04	0.15	0.09	0.04	0.05

Note. \hat{b} = estimated item difficulties; SE = standard error of parameter estimate. WM = white male; WF = white female; BM = black male; BF = black female; HM = Hispanic male; HF = Hispanic female.

IRT item bias statistics for the Mathematics Usage test are presented in Table 3. One unusual feature of this table is that there were only 4 items found to be biased in the white-male/Hispanic-male comparisons. There were 9, 16, and 15 items found to be biased in the cross-sex comparisons, and 10 items were biased in the white-male/black-male comparisons. These latter results are more typical of measurement equivalence studies (e.g., Lord, 1977) than are the white-male/Hispanic-male results. Another surprising aspect of Table 3 is that the direction of bias varies from item to item: Some items were easier for a group compared to white men (\hat{b}_i for white men > \hat{b}_i for the comparison group) and some items were harder for the comparison group (\hat{b}_i for white men < \hat{b}_i for the comparison group). We had expected that biased items would generally be biased in favor of white men.

Test-characteristic curve values, which give the expected number-right score as a function of the latent trait, are presented in Table 4 for the Mathematics Usage test. (Means, standard deviations, and reliabilities of the number-right test scores

are given in Table 9.) Table 4 should reveal any cumulative measurement bias against a group, because only unbiased items were used to link metrics. The differences in expected number-right scores are surprisingly small: Individuals from each of the other five groups always have expected number-right scores within 1 point (on a 40-point scale) of the expected number-right score of a white man with the same ability. Because any cumulative measurement bias for or against one group would elevate or depress their test-characteristic curve, we must conclude that there is no evidence of cumulative bias on the Mathematics Usage test and no revision of the test would be necessary. This clearly contradicts the conclusion indicated by inspecting proportions of correct answers.

The differences between the expected number-right scores in Table 4 are small for two reasons. First, the direction of bias is inconsistent across items and so cancellation occurs. Second, despite some very large chi-square statistics in Table 3 (e.g., 79 for the white-male/black-female comparison on Item 13), the differences in estimated item difficulties are not especially large (e.g., $\hat{b}_{13} = -0.07$ for white men and $\hat{b}_{13} = -0.46$ for black women; see Table 5).

Methodological digression on cumulative measurement bias. The finding of no cumulative measurement bias on the Mathematics Usage test was a surprise. It lead several readers of an earlier draft of this article to suggest that some methodological artifact in the IRT analysis would always force observed test-characteristic curves of female and minority groups to match the white male curve *even when there really were differences*.

To check this hypothesis, two simulated data sets were created and subjected to the measurement bias analyses. The first data set consisted of 1,500 response vectors generated with normal (mean zero, unit variance) abilities and item parameters set equal to the parameter estimates obtained for the white men. The second data set also contained 1,500 response vectors. These vectors were generated with normal ($M = -.75$, unit variance) abilities and the same item parameters as the first data set, except that one third of the items were made to be biased by increasing their item difficulties by .5. The IRT item bias analysis was perfect: All 13 of the items simulated to be biased were classified as biased in the IRT analysis, and none of the 27 unbiased items was classified as biased. Furthermore, each of

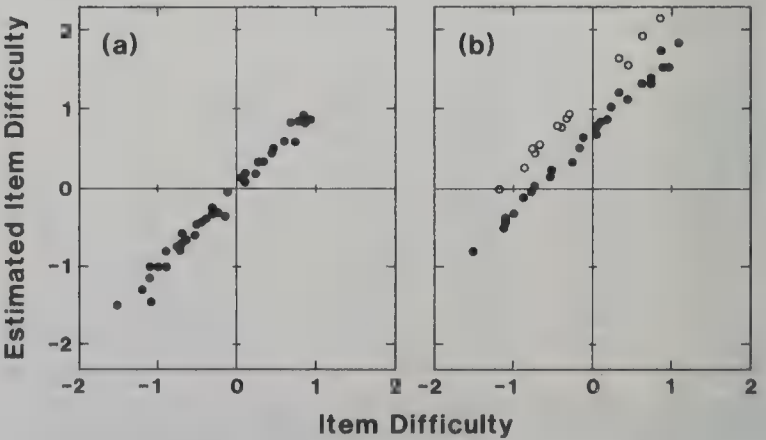


Figure 2. Estimated item difficulties plotted against the white male estimated item difficulties.

Table 6
Eigenvalues of Reduced Tetrachoric Correlation Matrices for the ACT English Usage Test

Sample	n	Eigenvalue no.						
		1	2	3	4	5	6	7
Mixed (men and women)	3,000	16.85	2.67	1.95	1.33	1.09	1.00	0.80
White men	1,500	15.26	2.52	1.95	1.62	1.28	0.98	0.91
White women	1,500	16.33	2.87	2.16	1.56	1.13	1.05	0.96
Black men	1,500	14.46	3.23	2.37	1.44	1.13	1.06	0.95
Black women	1,500	14.09	3.40	2.30	1.50	1.06	1.02	0.99
Hispanic men	1,441	15.94	3.11	1.88	1.44	1.22	1.07	0.91
Hispanic women	1,500	16.19	2.92	1.89	1.43	1.36	1.04	0.90
Harrison's	1,000	16.97	3.26	2.61	2.41	2.20	1.61	1.47

Note. ACT = American College Testing Assessment. Largest off-diagonal correlations were used as communality estimates for the 75-item ACT test and the simulated 70-item test created by Harrison (1986).

the two observed test-characteristic curves for the two groups was close to the corresponding test-characteristic curve computed from the simulation parameters. This shows that the IRT analysis can detect cumulative measurement bias when one third of the items are biased. The estimated item difficulty parameters for the two simulated data sets are plotted against the white male item parameter estimates in Figure 2. The items simulated to be biased in the second data set are indicated by open circles in Figure 2b and are obvious outliers.

Substantive content of biased Mathematics Usage items. Estimated item difficulties (\hat{b} s) and standard errors for the ACT Mathematics Usage items found to be biased in three or more comparisons are given in Table 5. (Notice that neither Item 4 nor Item 28, which showed bad fits in Figure 1, are among the items in Table 5.) With the exception of two comparisons (both were with black males), Items 14, 21, and 36 are uniformly easier for white males than for the other five groups and Items 1,

9, 13, 16, 22, 26, 30, and 38 are uniformly harder for white males. These latter eight items are calculation problems. Item 1, for example, presents two equations in two unknowns and asks the examinee to compute the solution. Item 9 asks the examinee to add four signed digits, two of which were placed within absolute value signs. The other six items biased against white males are similar in that they all involve no more than calculations. Item 14, in contrast, requires the examinee to compute the probability (assuming random sampling) of the union of two events from a table that lists population frequencies, and Item 21 concerns three individuals sharing money in certain ratios. Item 36, unfortunately, does not fit into the pattern that emerged from the eight items biased against white males and the other two items biased for white males; it requires calculating the difference of the product of three signed digits and their sum. Thus, it is a calculation problem that is similar to the items biased against white males.

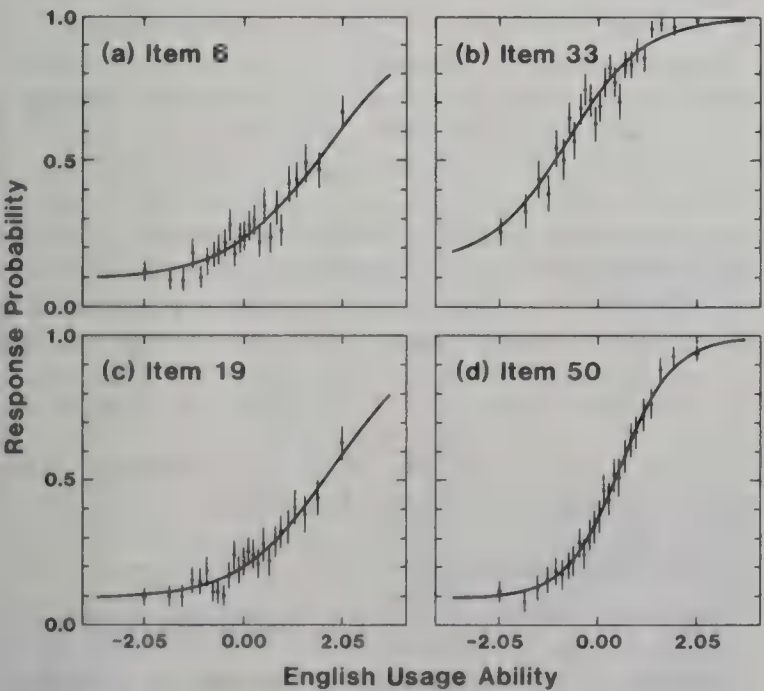


Figure 3. Empirical proportions, confidence intervals, and estimated item-characteristic curves for four English Usage items.

English Usage Test

Table 6 presents the eigenvalues of the reduced matrix of tetrachoric correlations for the 75-item English Usage test. These eigenvalue patterns do not seem to be comparable to the eigenvalues presented by Drasgow and Lissak (1983) because the eigenvalues in Table 6 are roughly 1.5 to 2.0 times larger than the eigenvalues previously published. Consequently, the eigenvalues were obtained from a data set created by Harrison (1986) to simulate eight group factors and a moderately prepotent general factor underlying a 70-item test. Harrison found that the item parameter estimates computed by LOGIST for samples of $n = 1,000$ were very accurate under these conditions. The patterns of eigenvalues in Table 6 for the English Usage test are all similar to the eigenvalue pattern from Harrison's data. Thus, we would expect accurate item parameter estimates despite violations of the unidimensionality assumption of the three-parameter logistic model. (Although Harrison's item responses were generated to simulate items where guessing the answer was not possible, Drasgow and Lissak's results show how to compare the eigenvalues from the English Usage test to Harrison's eigenvalues. They found that guessing reduces the size of the first eigenvalue but has smaller effects on the second and

Table 7
Chi-Square Statistics for English Usage Items Found to Be Biased in Comparisons With White Men

Group							Group						
Item	WF	BM	BF	HM	HF	Total	Item	WF	BM	BF	HM	HF	Total
1						0	39						0
2			26 ^h			1	40						0
3						0	41						0
4						0	42						0
5						0	43					19 ^h	1
6						0	44		17 ^h	20 ^h			2
7					16 ^e	1	45	18 ^h					1
8			18 ^e			1	46						0
9						0	47						0
10					26 ^e	1	48			17 ^h			1
11						0	49						0
12			20 ^e			1	50						0
13			27 ^h			1	51						0
14						0	52		21 ^e	34 ^e			2
15				23 ^e	24 ^e	2	53			30 ^e		18 ^e	2
16						0	54			22 ^e			1
17		18 ^h	26 ^h	16 ^h		3	55						0
18						0	56	55 ^h				31 ^h	2
19						0	57						0
20		24 ^h	50 ^h			2	58						0
21			21 ^h			1	59						0
22			23 ^e		17 ^e	2	60	21 ^e	26 ^e	30 ^e		21 ^e	4
23						0	61						0
24		15 ^e				1	62					17 ^e	1
25	16 ^h					1	63		25 ^h	25 ^h			2
26						0	64	30 ^h					1
27		31 ^h	33 ^h	19 ^h		3	65						0
28				16 ^e		1	66	29 ^h		20 ^h		28 ^h	3
29						0	67						0
30						0	68			17 ^e		23 ^e	2
31						0	69						0
32		20 ^e	40 ^e		20 ^e	3	70						0
33			16 ^e			1	71						0
34						0	72						0
35						0	73						0
36						0	74					19 ^e	1
37						0	75	17 ^e	16 ^e	20 ^e			3
38			22 ^e			1							
Number significant within group:							—	7	10	22	4	13	—

Note. WF = white female; BM = black male; BF = black female; HM = Hispanic male; HF = Hispanic female. ^e = easier (\hat{b}_i for indicated group is less than b_i for white men); ^h = harder (\hat{b}_i for indicated group is larger than b_i for white men). Nonsignificant statistics are omitted.

subsequent eigenvalues. This is the pattern found in Table 6 for the English Usage data sets.)

The fits of the three-parameter logistic item-characteristic curves to data for several items are shown in Figure 3. Figure 3a and Figure 3b again present the results for items that were badly fitted, and Figure 3c and Figure 3d show results that are typical of the vast majority of items. As with the Mathematics Usage test, the fits seem adequate for us to assess measurement equivalence.

The results of testing for item bias by Lord's chi-square test are shown in Table 7. Notice that only in the black female comparison is there a very large percentage (29%) of biased items. Comparisons of the white men to the other four groups show that roughly 5% to 17% of the English Usage items are biased.

Test-characteristic curve values for the English Usage test are presented in Table 8. The expected number-right scores are

again remarkably similar across the six groups: The largest difference between expected number-right scores is 1.0 point.

Finally, the linking constants for the two tests are presented in Table 9. These linking constants would be used in the equation

$$\text{Rescaled } \theta = A(\theta \text{ in original, standardized form}) + B.$$

The transformation from, say, the white female metric to the white male metric for the English Usage test is

$$(\theta_{WF} \text{ in white male metric}) = 1.00$$
$$\times (\theta_{WF} \text{ in white female metric}) + 0.22.$$

As a rough check on the reasonableness of the iterative Stocking and Lord (1983) linking constants, we derived analogous coefficients that could be used to transform test scores ex-

Table 8
Expected Number-Right Score as a Function of Ability
for the English Usage Test

Ability	Group					
	WM	WF	BM	BF	HM	HF
-2.5	19.3	19.5	20.0	20.3	19.4	19.7
-2.0	22.7	22.7	23.2	23.5	22.8	23.1
-1.5	27.0	26.8	27.2	27.5	27.2	27.4
-1.0	32.4	32.0	32.3	32.5	32.6	32.8
-0.5	38.8	38.4	38.5	38.7	39.0	39.2
0.0	46.0	45.6	45.6	45.6	46.0	46.4
0.5	52.9	52.7	52.6	52.4	52.8	53.3
1.0	58.7	58.7	58.8	58.4	58.6	59.3
1.5	63.4	63.5	63.8	63.2	63.4	64.0
2.0	67.0	67.1	67.6	66.9	67.0	67.5
2.5	69.6	69.7	70.1	69.4	69.6	70.0

Note. All ability metrics were transformed to the white male metric, which was scaled to have a mean of zero and a standard deviation of one. WM = white male; WF = white female; GM = black male; BF = black female; HM = Hispanic male; HF = Hispanic female.

pressed in, say, the white women’s standard score metric into the white men’s standard score metric; that is,

(z_{WF} in white male metric)

$$= \tilde{A}(z_{WF} \text{ in white female metric}) + \tilde{B}.$$

It is easy to show that

$$\tilde{A} = SD_{WF}/SD_{WM}$$

and

$$\tilde{B} = (M_{WF} - M_{WM})/SD_{WM}.$$

The crude linking coefficients \tilde{A} and \tilde{B} agree surprisingly well with the iterative Stocking and Lord (1983) coefficients, particularly on the English Usage test. The \tilde{B} s also agree with the B s on the Mathematics Usage test, but for this test the \tilde{A} s are all smaller than the A s. This is probably due to a floor effect: Number-right scores are bounded from below by 0 (and, due to guessing, are actually unlikely to be less than 5 or 6), whereas θ s are unbounded. Therefore, it is not surprising that the Mathematics Usage standard deviations of number-right scores for black men and women are relatively small. These small standard deviations in the number-right scale cause the \tilde{A} s to be small. In contrast, because the θ scale does not have a floor effect the standard deviations of $\hat{\theta}$ for the black men and women will not be reduced, and their A s will tend to be larger than the \tilde{A} s.

Discussion

Our results revealed little evidence of cumulative measurement bias when appropriate methods were used. Items were biased in different directions, and so their effects on total test scores tended to cancel. In addition, the effect sizes were usually small. Therefore, with all due consideration to methodological limitations, we must conclude that the ACT Assessment English Usage and Mathematics Usage tests provide equivalent measurement for Hispanic, black, and white men and women.

The finding of equivalent measurement for the ACT tests underscores the problems that are likely to occur when tests must be constructed according to court-mandated principles rather than the best available measurement theory: The statistical approach endorsed in the Golden Rule and *Allen v. Alabama* cases can yield very misleading information about the measurement properties of a test.

We cannot make the stronger conclusion that individual items provide equivalent measurement. Even though several steps were taken to minimize Type I errors (a stringent alpha level was used, ability metrics were linked iteratively, the fit of the model to the data was checked), we found many items to be biased. It should be emphasized that the use of very large samples allowed us to detect even minor differences between groups. Nonetheless, it seems appropriate to conclude that the varied experiences of the different racial and sex groups are translated into somewhat divergent patterns of responses to particular items.

In IRT item bias studies there has been little concern with the distinction between *significant* differences and *practically important* differences; almost all emphasis has been on statistical significance. The results presented in Tables 4 and 8 show that it is important to go beyond significance and consider effect size. Because most uses of tests are based on total test scores, differences in test-characteristic curves seem to be the natural quantity to index effect size. We conducted several simple analyses that determined the effect sizes that result when all biased items are biased in the same direction. These effect sizes were surprisingly small. For example, the Results section of this article described a simulation study in which 13 items (of 40) were created to be biased by increasing their item difficulty parameters by .5. The differences in test-characteristic curves for the unbiased and biased versions of the test never exceeded 3 points (about 0.4 SD) on the 40-point number-right scale. Thus, a

Table 9
Means, Standard Deviations, Reliabilities, and
Linking Constants for the Mathematics
Usage and English Usage Tests

Group	<i>M</i>	<i>SD</i>	Reliability	<i>A</i>	<i>B</i>	\tilde{A}	\tilde{B}
Mathematics Usage test							
WM	23.9	8.7	.91	— ^a	— ^a	— ^a	— ^a
WF	20.0	8.3	.89	1.00	−0.47	0.95	−0.45
BM	14.8	7.4	.86	1.07	−1.19	0.85	−1.05
BF	13.6	7.1	.85	1.09	−1.28	0.82	−1.18
HM	19.8	8.6	.90	1.03	−0.47	0.99	−0.47
HF	16.5	8.0	.88	1.06	−0.97	0.92	−0.85
English Usage test							
WM	45.8	12.2	.90	— ^a	— ^a	— ^a	— ^a
WF	48.0	12.2	.91	1.00	0.22	1.00	0.18
BM	35.2	12.3	.90	1.13	−0.89	1.01	−0.87
BF	37.7	12.0	.90	1.02	−0.67	0.98	−0.66
HM	40.4	12.9	.91	1.08	−0.44	1.10	−0.44
HF	41.8	12.8	.91	1.05	−0.35	1.05	−0.33

Note. WM = white male; WF = white female; BM = black male; BF = black female; HM = Hispanic male; HF = Hispanic female.

^a $A = \tilde{A} = 1.00$, and $B = \tilde{B} = 0.00$, by definition for the white men.

difference in item difficulties of 0.5, which ordinarily results in a very large and significant chi-square statistic, nonetheless has only a moderate effect on the test-characteristic curve even when one third of the items are biased. In sum, when effect sizes are considered, the similarities in measurement properties shown in Tables 4 and 8 across sex and racial groupings seem more salient than the differences.

We had originally expected to find more biased items on the English Usage test than on the Mathematics Usage test because mathematical items have a verifiable answer. English usage, in contrast, is a matter of convention, which can differ across cultural groups. Any differences in conventions concerning grammatical structures from those of white men would produce item bias. This prediction about the relative frequency of biased items on the two tests was incorrect. Two post hoc explanations for this finding are available. First, white men may be taught mathematics in ways that are somewhat different than in the other groups. This hypothesis is suggested by the 14 items found to be biased in at least two of the three cross-sex comparisons: Nine of these items were easier for all three female groups, and the other five items were uniformly harder for the women. Of the items that are easier for women, eight involve direct calculations. The other problem contains a paragraph that verbally tells the examinee what to do and then requires a simple calculation. The items that are more difficult for women include two word problems, a table-reading problem, a pie-chart problem, and a calculation problem. With the sole exception of this last item, these results would be expected if calculations were particularly emphasized in the mathematical educations of women, and more analytical topics were emphasized for men.

The other explanation of the greater relative frequency of biased items on the Mathematics Usage test than on the English Usage test is methodological in nature. In the statistical theory that underlies Lord's chi-square item bias statistic it is assumed that the person parameters are known rather than estimated. The use of estimated person parameters in place of the true values tends to inflate the chi-square statistic (McLaughlin, 1986). Person parameter estimates from a test that is sufficiently long and informative should be almost identical to the true values, and so the assumption about the person parameters will be approximately satisfied. In this case we would not expect the chi-square statistic to be inflated. Thus, our finding of relatively less bias on the 75-item English Usage test may have resulted at least in part because it is longer than the 40-item Mathematics Usage test.

This methodological difficulty with "conditional" maximum likelihood estimation emphasizes the importance of "marginal" maximum likelihood estimation methods. In this latter approach to estimation the person parameters are removed from the likelihood equation by integrating with respect to an ability distribution, which eliminates the necessity of assuming that the person parameters are known. In our attitude measurement research we have found fewer biased items and more interpretable results when marginal maximum likelihood estimation methods were used to estimate parameters of the two-parameter logistic model. Work by Bock and his colleagues (Bock & Aitkin, 1981; Bock & Lieberman, 1970) on estimation for three-parameter models is a step in the right direction for measurement bias research on multiple-choice tests.

Conclusion

Measurement bias should not be investigated using the proportion-correct statistic because it confounds bias with between-group differences in the attribute measured by the test. Between-group differences are expected whenever the "environments" of groups differ, and it is clear that environments differ for men, women, whites, blacks, and Hispanics. In this study the proportion-correct statistic was shown to lead to an incorrect conclusion about measurement bias. To reach correct conclusions about bias it is necessary to use a method that does not confound measurement bias with between-group differences. One such method is provided by IRT. No evidence of cumulative measurement bias was found across six groups for the ACT Mathematics Usage and English Usage tests when the psychometrically appropriate analysis was conducted.

References

- Allen v. Alabama State Board of Education, No. 81-697-N (consent decree filed with United States District Court for the Middle District of Alabama Northern Division, 1985).
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-459.
- Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35, 179-197.
- Cordes, C. (1985, February). ETS to reweigh test items' racial bias. *APA Monitor*, pp. 26, 28.
- Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin*, 95, 134-135.
- Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology*, 68, 363-373.
- Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, 7, 189-199.
- Dunbar, S. B., & Novick, M. R. (1985). *On predicting success in training for males and females: Marine Corps clerical specialties and ASVAB Forms 6 and 7* (Tech. Rep. No. 85-2). (Available from CADA Research Group, 356 Lindquist Center, University of Iowa, Iowa City, IA 52242).
- Golden Rule Insurance Company et al. v. Washburn et al., No. 419-76 (stipulation for dismissal and order dismissing cause, Circuit Court of the Seventh Judicial Circuit, Sangamon County, IL, 1984).
- Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics*, 11, 91-115.
- Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.
- Hunter, J. E. (1975, December). *A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items*. Paper presented at the National Institute of Education Conference on Test Bias, Bethesda, MD.
- Jaschik, S. (1986). Critics and defenders of standardized tests weigh "truth-in-testing" bills in New York. *Chronicle of Higher Education*, 32, 13-14.
- Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research*, 43, 139-161.
- Linn, R. L. (1982). Ability testing: Individual differences, prediction

- and differential prediction. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, Consequences, and Controversies* (pp. 335-388). Washington, DC: National Academy Press.
- Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19-29). Amsterdam: Swets & Zeitlinger.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- McLaughlin, M. E. (1986). *Computing Lord's item bias statistic when abilities and item parameters are estimated simultaneously*. Master's thesis, Department of Psychology, University of Illinois at Urbana-Champaign.
- Petersen, N. S. (1980). Bias in the selection rule—bias in the test. In L. J. T. van der Kamp, W. F. Langerak, & D. N. M. de Gruijter (Eds.), *Psychometrics for educational debates* (pp. 103-122). New York: Wiley.
- Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, 13, 3-29.
- Pitcher, B. (1974). Predicting law school grades for female law students. In Law School Admission Council, *Annual council report* (pp. 555-575). Washington, DC: Law School Admission Council.
- Pitcher, B. (1975). A further study of predicting law school grades for female law students. In Law School Admission Council, *Annual council report* (pp. 107-150). Washington, DC: Law School Admission Council.
- Scheuneman, J. D. (1980). Latent-trait theory and item bias. In L. J. T. van der Kamp, W. F. Langerak, & D. N. M. de Gruijter (Eds.), *Psychometrics for educational debates* (pp. 139-151). New York: Wiley.
- Scheuneman, J. D. (1982). A posteriori analyses of biased items. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 180-198). Baltimore, MD: Johns Hopkins University Press.
- Segall, D. (1982). *Iterative linking of ability metrics*. Unpublished manuscript, University of Illinois at Urbana-Champaign, Department of Psychology.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.
- Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). *LOGIST—A computer program for estimating examinee ability and item characteristic curve parameters* (Research Memorandum No. 76-6). Princeton, NJ: Educational Testing Service.

Received March 4, 1986

Revision received June 23, 1986 ■

The Power of the Schmidt and Hunter Additive Model of Validity Generalization

Edward R. Kemery
Tulane University

Kevin W. Mossholder
Management Department, Auburn University

Lawrence Roth
Tulane University

Monte Carlo simulation and infinite sample-size analysis were used to test inferences based on the Schmidt and Hunter Additive Model of Validity Generalization when ρ in some instances was zero. Results of both large ($n = 1,000$) and small ($n < 100$) sample simulations, and corroborating evidence from infinite sample-size analysis, suggest that when ρ is zero in upward of 30% of the cases (and ρ is .6 for the remaining cases), the Schmidt and Hunter procedure could prompt erroneous inferences of generalizability.

Personnel psychologists have long noted that the correlation between abilities and job performance varies across situations (e.g., Ghiselli, 1966), even when the job and ability are well matched. Two explanations have been advanced to explain this occurrence. One, *validity generalization*, posits that variation in validities is due to statistical and methodological artifacts (e.g., Schmidt & Hunter, 1977). The alternative explanation, which Schmidt and Hunter (1977) labeled the *theory of situational specificity* and attributed to Ghiselli (1966), suggests that validity differences are caused by situational factors.

Validity Generalization

An important heuristic development in industrial psychology in the past decade has been the advancement of a theoretical model of validity generalization (e.g., Callender & Osburn, 1980; Schmidt & Hunter, 1977). The theory of validity generalization hypothesizes that the relation between an ability and job performance will be consistent across situations. Models of validity generalization have been developed to account for the common observation that validity coefficients derived from the same (or similar) test-job combination vary across situations. In fact, this variability has been so large that in some instances negative correlations have been found when positive ones were expected (Ghiselli, 1966).

The theory of validity generalization predicts that in reality there is no substantial variance in test validities. Proponents argue that the observed variance in validities can be accounted for by the following seven artifacts: (a) sampling error, (b) criterion unreliability, (c) predictor unreliability, (d) range restriction on the predictor, (e) criterion contamination and deficiency, (f)

clerical errors, and (g) imperfect construct validity of the predictor (Schmidt & Hunter, 1977).

The validity generalization hypothesis has been tested using a variant of meta-analysis (e.g., Pearlman, Schmidt, & Hunter, 1980). This procedure uses empirical data from multiple studies to estimate (a) the true relation between predictor and criterion constructs and (b) a confidence interval around this estimate. The general relation between predictors and criteria is estimated from a weighted average of observed validity coefficients after correcting them for attenuation due to range restriction and unreliability. The variance attributable to artifacts is then estimated and compared with the variance in observed validities. If the variance across studies expected from artifacts accounts for a substantial portion of the observed variance (i.e., $\geq 75\%$), and if the estimated population correlation coefficient is nonzero, validity generalization is inferred. Support for the validity generalization hypothesis is reported in several studies covering a wide variety of jobs (see Burke, 1984, for a review).

Situational Specificity

The concept of situational specificity refers to the contextual moderation of test-job validities. Ghiselli (1966) defined situational specificity as differences in the factor structure of job performance due to specific variables that determine the nature of performance in particular settings. Burke (1984, p. 94) defined it in terms of situational moderators, for example, between-job task or behavior differences that cause a predictor to be valid for some jobs, *but not for others*. Osburn, Callender, Greener, and Ashworth (1983) stated additionally that test validity may vary as a function of such variables as organizational differences, geographical location, and job content differences. Thus, the situational specificity notion suggests that test-performance relations vary across settings as a function of a host of contextual factors.

Situational specificity has been operationally defined by validity generalization researchers as the variance in validity coefficients remaining after removing variance expected from

We wish to thank Larry James, Richard Arvey, and several anonymous reviewers for their helpful suggestions on an earlier version of this article.

Correspondence concerning this article should be addressed to Edward R. Kemery, Department of Psychology, Tulane University, New Orleans, Louisiana 70118.

methodological artifacts. According to the Schmidt and Hunter (e.g., Schmidt, Hunter, & Pearlman, 1982) 75% rule, if measurement artifacts account for at least 75% of the variance in observed validities, the situational specificity hypothesis should be rejected in favor of validity generalization. Their operationalization does not directly assess situational specificity. Rather, its effect is inferred from what remains after removing the hypothesized influence of artifacts. This procedure is analogous to a hierarchical multiple regression analysis in which artifact variables are entered as predictors and residual variance is inspected for evidence of situational specificity.

Validity Generalization Versus Situational Specificity Versus Transportability

Although sometimes taken to be distinct, validity generalization and situational specificity are not necessarily mutually exclusive. Both generalizability and situational specificity can occur simultaneously. Validity generalization researchers have recognized this when suggesting ways to interpret Bayesian priors obtained from meta-analysis. For example, Pearlman et al. (1980) have argued that even when the validity generalization hypothesis (i.e., true variance = 0) cannot be rejected according to their 75% rule, validity generalization may still be warranted. They recommended calculating a 90% credibility value, which is defined as the 10th percentile validity value (Pearlman et al.). If the 90% credibility value is greater than zero, then validity generalization is supported.

It is apparent that the term *validity generalization* has been used in two different ways. As initially proposed, validity generalization meant that variations in validity could be attributed to artifacts (Schmidt & Hunter, 1977, p. 529), implying that the variance in true validities is zero. Situational specificity, however, is usually taken to mean that variation in true validities is nonzero. To avoid the confusion that could arise in instances where validity variance was modest and the average validity substantial (e.g., .6), for the purpose of the present study, we define cross-situational consistency as the condition (or hypothesis) that true validity variance is zero (James, Demaree, & Mulak, 1986). When true validity variance is nonzero, this will be referred to as *situational specificity*. The term *transportability* will be used to refer to situations in which cross-situational consistency is not supported, but the 90% credibility value is greater than zero.

Although previous researchers considering the cause of validity fluctuations have tended to advocate either validity generalization or situational specificity, it is conceivable that they may coexist. Granting this possibility, a reasonable alternative to the cross-situational consistency hypothesis would be that observed validity coefficients are co-determined by measurement factors and situational variables. Specifically, it is not unreasonable to postulate that the cause of some variance attributed to artifacts is in the situation itself. For example, validity generalization advocates argue that criterion reliability differences between job settings account for some of the variability in observed validities. Although this is undoubtedly true, the explanation for some reliability differences actually may be in the situations themselves. Using performance measures as an example, sources of criterion error may take many forms (Thorndike,

1949). Although situational factors are referred to in terms such as "changes in the testing environment" (Dunnette, 1966), "temporary reasons that would apply to almost any situation" (Guion, 1965), or "situation-induced error" (Lyman, 1978), the shared idea is that they contribute to reliability of performance measures, particularly in the case of subjective ratings.

The point that differences in situations may contribute to differences in measurement reliability is an important one because it would then follow that variance across situations due to reliability (and perhaps other measurement "artifacts"), and variance due to situational factors, are not orthogonal. If this is the case, when artifact variance is removed during meta-analysis, some of the variance due to situational factors is inadvertently removed. Thus, the meta-analytic procedures, as currently used, may be incapable of ruling out situational factors as alternative explanations of cross-situational validity differences.

A recognition that some variation in correlations might be real suggests that, in opposition to the usual psychometric assumption of an underlying unimodal validity distribution, there could be more than one subpopulation of validities (Algera, Jansen, Roe, & Vijn, 1984). Because there is no compelling substantive theoretical reason for the assumption of a unimodal distribution, it is perhaps just as reasonable to postulate a multimodal continuum of validities. One basis for this hypothesis involves the impact of the environment on jobholder behavior.

The concept of situational specificity by definition suggests that environmental variables are important determinants of job performance. Along this line, Peters and O'Connor (1980) have provided a taxonomy for describing these variables. Their framework includes variables such as (a) job-related information, (b) tools and equipment, (c) materials and supplies, (d) budgetary support, (e) help from others, (f) task preparation, (g) time availability, and (h) the physical work environment. They consider the presence or absence of these variables to be important in determining job performance, affective responses, and turnover. Empirical support for this contention has been provided by O'Connor et al. (1984). A condition of particular relevance to situational specificity is when these variables serve to limit worker productivity, that is, the concept of situational constraint.

Situational Constraints

A situational constraint is a variable in the work setting that has a negative impact on jobholder behavior (Peters & O'Connor, 1980). Components of the above-mentioned taxonomy are potential constraints because they are necessary for workers to realize full performance potential. Moreover, these situational factors may not affect each worker similarly. High-ability jobholders, because they have the capacity to perform extremely well, will be affected most (Schneider, 1975, 1978). In other words, situational constraints place a ceiling on the behavior of only those workers having the ability to perform above it. This particular Person \times Situation interaction would serve to attenuate an observed correlation between ability and performance relative to a lesser constrained situation, not because of a statistical artifact, but rather because the variance in performance

was actually restricted by the situation (Peters & O'Connor, 1980; Schmidt et al., 1985).

Because to some unknown degree, validity differences between settings could be the product of situational phenomena (e.g., situational constraints), in addition to measurement artifacts, one could postulate that a validity coefficient is produced by a unique configuration of contextual determinants and measurement error (cf. James et al., 1986). Further, because of numerous inhibitory constraints, it would not seem unreasonable to postulate situations where ability-job-performance correlations would be minimal, differing from zero in neither a statistical nor practical sense. This possibility has been recognized (Algera et al., 1984; Schmidt, Hunter, & Pearlman, 1981). It is believed, however, that validity generalization algorithms would rarely have the statistical power to detect it (Schmidt, Hunter, & Pearlman, 1981, p. 174).

Previous Validity Generalization Studies

Several validity generalization studies have used real-world data within meta-analytic frameworks (e.g., Schmidt, Hunter, & Caplan, 1981). In these approaches, the true underlying distributions of validity are unknown. Validity generalization researchers have implicitly assumed that validities are the same across situations and, within this context, have supported generalizability (i.e., cross-situational consistency or transportability) about 50% of the time (Schmidt et al., 1982). Concern about the accuracy of these inferences is warranted inasmuch as there is no standard against which to gauge their veracity. One means for circumventing this problem would be to use Monte Carlo simulation.

Callender and Osburn (1980, 1981) used simulation techniques to assess the accuracy of the Schmidt and Hunter (1977) model. Aside from suggesting minor modifications of the original Schmidt and Hunter algorithms, their work generally supported the validity generalization model. (See, however, Osburn et al., 1983, for a cautionary note.) A closer look at Callender and Osburn's (1980) method reveals that they also assumed that validities generalize. That is, although they allowed true validity (ρ) to vary from .1 to .9, they did not include a condition in which ρ was zero. Thus, the ability of validity generalization algorithms to detect such an extreme instance is unknown.

The present study breaks from the practice of assuming a single underlying population correlation coefficient. Rather, we propose that there are different kinds of situations, each with its own population validity. To simplify matters, we will assume that because of a systematic distribution of situational constraints, there is one kind of situation in which correlations of ability and performance are neither statistically nor practically different from zero. We further make the simplifying assumption that in all other situations, the population correlation coefficient is .6. Thus, within the context of this admittedly contrived dichotomous world, we asked how accurate the Schmidt and Hunter additive model algorithms are in detecting instances of zero validity? These algorithms were chosen because of their computational simplicity and because only trivial differences have been found when other algorithms were used instead (Schmidt et al., 1982).

Monte Carlo simulation and infinite sample-size analysis

(Callender & Osburn, 1980) provided a guidepost against which to gauge the accuracy of inferences from the Schmidt and Hunter procedure.

Method

Computer Program

The computer program was designed to generate data from either a unimodal distribution ($\rho = .6$) or a bimodal one ($\rho = 0$ and $\rho = .6$), based on proportions entered by the user. These two values of ρ were selected to simulate extreme instances of validity. Furthermore, we reasoned that in the real world, a predictor's use would be predicated on theoretical or empirical considerations. Thus, the likelihood that a predictor would be valid would probably exceed some arbitrary minimum. Therefore, we varied the percentage of time ρ was equal to .6 from 100% to 50%. (If the percentage at ρ of .6 was 70%, the percentage at ρ of zero was automatically 30%). When unimodality was specified, ρ was .6 for each simulated validity study. On the other hand, when bimodality was specified, the program first randomly selected the value of ρ from which to start.

Next, using algorithms supplied by Box and Muller (1958), the program generated pairs of observations for each simulated subject. This was accomplished by combining pseudorandom normal deviates with the selected parameter to generate true predictor and criterion scores for each subject, as per the previously selected validity coefficient. To simulate observed predictor and criterion scores measured with error, these true scores were combined with pseudorandom normal deviates (Dunlap, Jones, & Bittner, 1983) selected from the distributions of assumed predictor and criterion reliabilities (see Appendix Tables A-1 and A-2, given by Schmidt, Hunter, Pearlman, & Shane, 1979, p. 260).

To simulate range restriction effects, only those observations exceeding a preselected cutting score were retained in each study. Cutting scores were determined by varying the selection ratio for each set of observations based on the distribution outlined by Schmidt et al. (1979, p. 261), which is presented in Appendix Table A-3. To illustrate, if a study had a selection ratio of one, and a sample size of 100, all pairs of observations were retained. However, if the selection ratio was .5, and 100 pairs of observations (i.e., subjects) were needed, 200 pairs of scores were generated with only the top 100 predictor (and corresponding criterion) scores retained.

Summarizing, the value of ρ for each simulated validity study was sampled randomly from the specified proportions at $\rho = .6$ and $\rho = 0$. Once the value of ρ for a specific validity study was selected, values of predictor and criterion reliability, and range restriction, were sampled randomly from the distributions given by Schmidt et al. (1979).¹ Pairs of scores for each simulated subject were generated and the observed correlation coefficient computed. Additionally, observed test and criterion reliabilities, and the ratio of restricted to unrestricted predictor variance, were calculated. Thus, we were able to use actual observed (rather than assumed) artifact values to estimate validity generalization.

The accuracy of our data-generating procedures was tested by conducting 1,600 simulated validity studies, using sample sizes (ranging from 29 to 321) selected to mirror the distribution obtained from a review of *Personnel Psychology's* Validity Information Exchange extending from 1959 to 1966. Results are shown in Table 1. Of interest are the average values of the artifacts. In every instance, they are extremely close to the input parameters. Also provided in Table 1 are the

¹ For Monte Carlo purposes, all levels of criterion reliability were crossed with $\rho = 0$. Note that if situational constraints are operating, the criterion reliability tends to be reduced proportionally to the degree of constraint.

Table 1
Means, Standard Deviations, and Intercorrelations Between Artifact Parameters, Retained Sample Sizes, and Observed Validity Coefficients

Study parameters	M	SD	1	2	3	4	5	6
1. Proportion of rho at zero	0.196	0.175	—	.0407	.0141	.0070	-.0089	-.2522
2. Range restriction	0.593	0.117	—	—	.0823	.0259	.0320	.1970
3. Predictor reliability	0.801	0.085	—	—	—	.0109	-.0195	.0934
4. Criterion reliability	0.601	0.146	—	—	—	—	.0201	.1236
5. Sample size	81.780	66.660	—	—	—	—	—	.0259
6. Observed correlation	0.200	0.172	—	—	—	—	—	—

intercorrelations of the artifacts and observed correlations. These intercorrelations were negligible, ranging from -.0195 to .0823, demonstrating the independence of the input values.

The relations between the observed correlations and artifact values presented in Table 1 are also in the expected direction. The percentage of time a rho of zero occurred was negatively related to the observed validity coefficients, whereas the selection ratio and predictor and criterion reliabilities were positively related to them. Sample size was uncorrelated with observed validity. Thus, the statistics generated from the program closely approximated specified distributions of input parameters.

The next steps in the program involved estimating the characteristics of a Bayesian prior using the procedure outlined by Pearlman et al. (1980). Specifically, the program estimated (a) the variance due to differences in criterion and test reliabilities, (b) the variance due to differences between studies in range restriction, (c) the variance due to sampling error, (d) predicted and residual variance, (e) the mean and standard deviation of "true" validities, (f) the proportion of observed variance due to artifacts, and (g) the 90% credibility value. The program also provided a relative frequency histogram of observed and true correlations for each simulated study.

Along with the Monte Carlo simulation, an infinite sample-size analysis (Callender & Osburn, 1980) provided expected values of Bayesian prior characteristics for each set of rhos. The mean, standard deviation, and 90% credibility values were calculated based on formulae for discrete random variables (McClave & Benson, 1982). These estimates were used to cross check the accuracy of our Monte Carlo results.

The Studies

Two meta-studies were undertaken to assess the accuracy of the Schmidt and Hunter procedure. The first meta-study was designed to simulate sample sizes usually found in the real world. These sample sizes were based on the distribution obtained from our review of the Validity Information Exchange extending from 1959 to 1966. The average sample size was 90, and *ns* ranged from 29 to 321. Obtained statistics from 800 individual studies, 100 for each of eight combinations of rho = .6 and rho = 0 listed in Table 2, were used to estimate the accuracy of the Schmidt and Hunter estimation procedure. Eight separate meta-analyses, one for each combination of rhos, were conducted within each meta-study.

The second meta-study was identical to the small-sample study, except that much larger samples (*N* = 1,000) were used. This analysis was conducted to determine if sample size influences validity generalization estimates. Note that in each study, the reported sample size refers to the actual number of subjects retained (i.e., only those above the cutting score specified by range restriction). Thus, actual sample sizes were much larger than reported.

Results

The pattern of observed validities obtained when bimodality was specified generally resembled the distribution of real-world validities reported by Ghiselli (1966). This was true in both the small- and large-sample meta-studies. An example of a distribution obtained from one of our studies is displayed in Figure 1. This particular pattern was generated from a small-sample run when the proportion of rho at zero was 10% and the proportion of rho at .6 was 90%.

Results obtained for the small-sample condition are shown in Table 2. In every case, the Schmidt and Hunter additive procedure estimated rho to be greater than zero. When true rho was always .6, the Schmidt and Hunter procedure provided a nearly accurate estimate (.5935). However, when the variance was actually zero, the Schmidt and Hunter algorithms estimated it to be less than zero. Overall, the proportion of observed variance accounted for by the estimated artifacts ranged from 0.449 to 1.537. When rho was zero as much as 10% of the time, cross-situational consistency would have been inferred.

Of special interest are results concerning transportability. When the proportion of variance explained by artifacts is less than 75%, Schmidt and his colleagues (e.g., Pearlman et al., 1980) maintain that validity may be transportable. Using the 90% credibility value, our small sample meta-study results sup-

Table 2
Results of Small-Sample Validity Generalization Meta-Study

Percentage ^a		rho ^b	<i>s</i> ^{2c}	90% cv	<i>s</i> ² explained ^d
.6	0				
100.0	0.0	.5935	<0	.5935	1.303
97.5	2.5	.6045	<0	.6045	1.537
95.0	5.0	.6125	<0	.6125	1.343
90.0	10.0	.5588	.0042	.3522	0.794
80.0	20.0	.4367	.0137	.0849	0.493
70.0	30.0	.4488	.0100	.1287	0.602
60.0	40.0	.3790	.0157	-.0176	0.449
50.0	50.0	.2722	.0124	-.0923	0.466

Note. cv = credibility value.
^a Percentage of the studies in which true rho was .6 and 0, respectively.
^b Estimate of rho based on the Schmidt and Hunter additive model.
^c Estimate based on the Schmidt and Hunter additive model (observed variance minus artifactual variance).
^d Proportion of observed variance accounted for by artifacts.

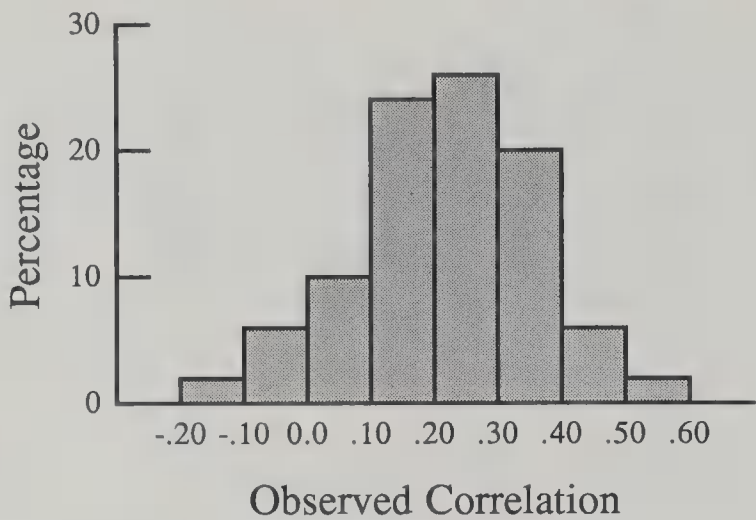


Figure 1. Frequency distribution of observed correlations when the proportion at rho = .6 was 90% and the proportion at rho = 0 was 10%.

port inferences of transportability until the percentage of rho at zero exceeds 30%!

Results for the large sample meta-study are reported in Table 3. The pattern of findings are similar to the small-sample meta-study, even though the large-sample estimates were each based on 1,000 retained observations. When rho was always .6, the Schmidt and Hunter procedure yielded a reasonable estimate (0.6127). When rho was zero 2.5% of the time, true validity variance was estimated to be less than zero. Even when rho was zero 10% of the time, based on the 75% rule, the cross-situational hypothesis would have been supported.

Based on an inspection of the 90% credibility value, an inference of transportability would have been supported when rho was zero 40% of the time. In other words, when zero validity existed in 40% of the population of validities, the Schmidt and Hunter algorithms would have prompted erroneous inferences of transportability.

Infinite Sample Size Analysis

Because our computer program was designed to sample from a bimodal world according to specified proportions, rho may be

Table 3
Results of Large-Sample Validity Generalization Meta-Study

Percentage ^a		rho ^b	s ^{2c}	90% cv	s ² explained ^d
.6	0				
100.0	0.0	.6127	<0	.6127	1.955
97.5	2.5	.5909	<0	.5909	1.180
95.0	5.0	.5739	.0017	.4459	0.815
90.0	10.0	.5736	.0011	.4717	0.857
80.0	20.0	.4574	.0111	.1379	0.260
70.0	30.0	.4525	.0132	.0877	0.254
60.0	40.0	.4040	.0145	.0361	0.208
50.0	50.0	.2198	.0116	-.1424	0.117

Note. cv = credibility value.
^a Percentage of the studies in which true rho was .6 and 0, respectively.
^b Estimate of rho based on the Schmidt and Hunter additive model.
^c Estimate based on the Schmidt and Hunter additive model.
^d Proportion of observed variance accounted for by artifacts.

Table 4
Results of Infinite Sample-Size Analysis

Percentage ^a		rho _{est}	s ²	90% cv
.6	0			
100.0	0.0	.6000	.0000	.6000
97.5	2.5	.5850	.0088	.4651
95.0	5.0	.5700	.0171	.4021
90.0	10.0	.5400	.0324	.3096
80.0	20.0	.4800	.0576	.1728
70.0	30.0	.4200	.0756	.0681
60.0	40.0	.3600	.0864	-.0162
50.0	50.0	.3000	.0900	-.0840

Note. cv = credibility value.
^a Percentage of the studies in which true rho was .6 and 0, respectively.

considered a discrete random variable. Therefore, the expected value of the mean and variance of the distributions (i.e., the Bayesian priors), given the relative proportions of rhos previously investigated, may be calculated precisely using formulae for calculating the mean and variance of discrete random variables (McClave & Benson, 1982).

Assuming that all variables are measured without error, expected values of the mean and standard deviation for each distribution may be calculated in the following manner: The mean is given by

$$pr(.6)*.6 = E(r), \tag{1}$$

where $pr(.6)$ refers to the proportion of times rho was .6, and $E(r)$ refers to the expected value of the mean of the distribution. The variance of the distribution is calculated by

$$\{pr(.6)*[.6 - E(r)]^2\} + \{pr(0)*[-E(r)]^2\} = var, \tag{2}$$

where $pr(0)$ is the proportion of times rho equals zero, and the remaining terms are as in Equation 1. To illustrate, when $pr(.6) = .8$, and $pr(0) = .2$, the expected value of the mean of the distribution = .48. The variance of this distribution = $.8*(.12)^2 + .2*(-.48)^2$, or 0.0576.

Because the 90% credibility value (CV) is defined as the 10th percentile validity value, it lies 1.28 SD below the mean (assuming a normal distribution). Therefore, the 90% credibility value is calculated by finding the validity value associated with a z score of -1.28. Thus, the 90% credibility value is found by

$$(-1.28 * var^{1/2}) + E(r) = 90\% CV. \tag{3}$$

In the present example, the 90% cv equals 0.1728.

The expected values for the distributions based on combinations of rhos used in the present study can be found in Table 4. Of particular interest is that in support of the Monte Carlo evidence, the proportion of rhos equaling zero would have to exceed 30% for transportability to not be inferred using the Schmidt and Hunter approach.

Discussion

The results of the meta-studies and infinite sample-size analysis suggest that the Schmidt and Hunter algorithms may

prompt erroneous inferences regarding the transportability of a predictor. Specifically, when it is known that a test is not valid upward of 30% of the time (within the specified parameters of this study), the Schmidt and Hunter formulae will still suggest transportability.

The implications of our findings need to be tempered by one caveat. We assumed, for the sake of simplicity, a bimodal world. Then, within this particular context, we asked the question: What if? It is perhaps just as reasonable to postulate more complex situations (see James et al., 1986). Thus, we are not arguing that bimodality is the true state of nature.

Obviously, the bimodal distributions chosen as the basis for this study represent an extreme situation and thus limit the generalizability of the present findings. Consideration of other less extreme distributions may be of value in determining how serious a threat the present findings are to validity generalization implications.

Given the framework of the present study, the accuracy of the Schmidt and Hunter additive model may be judged by comparing the Monte Carlo results with the expected values from the infinite sample-size analysis. Estimated ρ tended to be very accurate. However, variance estimates based on the Schmidt and Hunter procedure tended to be biased in the direction of overestimation, resulting in negative estimates when the true variance was negligible. This negative bias was expected and has been reported in previous research (Callender & Osburn, 1980). However, more recently developed validity generalization estimation procedures (e.g., Callender & Osburn, 1980) correct for this bias.

The 90% credibility values generally indicated, as corroborated by the infinite sample-size analysis, that transportability will be inferred until the proportion of times ρ equals zero exceeds 30%. Thus, the Type II error rate (i.e., inferring generalizability when one should not) for transportability would be in the neighborhood of .3, given the parameters of the present study.

Note that in the Monte Carlo studies, a slight degree of non-monotonicity was found between s^2 explained, estimated ρ , the 90% credibility value, and increasing degrees of induced specificity. That is, these estimates did not systematically increase when the percentage of ρ at zero decreased. The likely explanation is that this occurred because of sampling error fluctuations.

The finding that the proportion of cases in which ρ was zero must exceed 30% before transportability will not be inferred was surprising. One way to minimize the occurrence of a Type II error in this instance would be to adopt more stringent criteria for inferring transportability. The present study was not designed specifically to assess the 75% rule proposed by Schmidt et al., 1982. However, the results presented in Tables 2 and 3 suggest that, for the parameters used in our studies, a criterion of 85% might be more appropriate. With an 85% rule, the cross-situational consistency hypothesis would have been correctly rejected in 10 of 14 cases. Future work should investigate the adequacy of the 75% rule under a broader range of conditions (James et al., 1986; Osburn et al., 1983).

In five instances, the Schmidt and Hunter estimation procedures accounted for more validity variance than was actually present, in one instance by nearly 200%. A similar finding has

been reported using real-world data (Schmidt & Hunter, 1984). Although this might seem peculiar, a likely reason for these findings is that sampling error variance is overestimated by the algorithm used by Schmidt and his colleagues. Thus, when actual validity variance is small, even slight overestimations of sampling error variance could result in more than 100% of the observed variance being explained by artifacts (James et al., 1986).

It is clear that our findings are somewhat at odds with results from a number of recent validity generalization studies (Burke, 1984). Because our computer simulated results, as well as those of Sackett, Harris, and Orr (1986), caution against uncritical acceptance of validity generalization precepts (e.g., Schmidt & Hunter, 1981; Schmidt et al., 1985), it is suggested that more conceptualizing and research be conducted. For example, two studies viewed within the context of the present study, highlight some concerns regarding validity generalization results.

Schmidt, Hunter, and Pearlman (1981) investigated generalizability of general mental abilities tests in two different studies. In the first, either cross-situational consistency or transportability was inferred for cognitive abilities tests for a wide range of clerical occupations. Similarly, the results of their second study, using data obtained from 35 disparate Army jobs, indicated that several cognitive abilities were valid for all jobs studied. Although this study included only selected Army jobs, omitting critical combat jobs such as infantryman, tank crewman, and artilleryman, Schmidt et al. (1981) interpreted the results in conjunction with their previous studies of clerical workers to mean that aptitude tests are valid for all jobs. That is, they have stated categorically that the hypothesis of situational specificity is false where cognitive abilities are concerned.

The implications of their conclusions are far reaching, affecting such practices as job analysis and traditional approaches to validating tests (Schmidt & Hunter, 1981). Moreover, at least one researcher outside of the area of personnel psychology (Jensen, 1984) has chosen to interpret validity generalization results as support for the construct validity of general intelligence. However, in view of the findings presented here, the conclusions offered by Schmidt et al. would have been the same, even if for a meaningful proportion of the jobs there was no true relation between abilities and criteria. Thus, the conclusions offered by Schmidt and his co-workers (e.g., Schmidt & Hunter, 1981, 1984; Schmidt, Hunter, & Pearlman, 1981) regarding the omnibus validity of cognitive ability tests could be overly optimistic.

The major implication of the results obtained from our investigation is that the "problem" of validity generalization has not yet been solved. It appears that the Schmidt and Hunter model (and algorithms) will not permit accurate conclusions when it is assumed that validities are bimodal (and validity is either 0 or .6) rather than situationally consistent. Thus, scientific prudence is advocated when interpreting the results of validity generalization analyses. We concur with several authors (Algera et al., 1984; Burke, 1984; James et al., 1985) that much more work in this area is necessary before any overall conclusions can be offered.

References

- Algera, J. A., Jansen, P. G. W., Roe, R. A., & Vijn, P. (1984). Validity generalization: Some critical remarks on the Schmidt and Hunter procedure. *Journal of Occupational Psychology*, 57, 197-210.

- Box, G. E. P., & Muller, M. E. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, 29, 610-611.
- Burke, M. J. (1984). Validity generalization: A review and critique of the correlational model. *Personnel Psychology*, 37, 93-115.
- Callender, J. C., & Osburn, H. G. (1980). Development and test of a new model for validity generalization. *Journal of Applied Psychology*, 65, 543-558.
- Callender, J. C., & Osburn, H. G. (1981). Testing the constancy of validity with computer generated sampling distributions of the multiplicative model variance estimate: Results for petroleum industry validation research. *Journal of Applied Psychology*, 66, 274-281.
- Dunlap, W. P., Jones, M. B., & Bittner, A. (1983). Average correlations versus correlated averages. *Bulletin of the Psychonomic Society*, 21, 213-216.
- Dunnette, M. D. (1966). *Personnel selection and placement*. London: Brooks/Cole.
- Ghiselli, E. E. (1966). *The validity of occupational aptitude tests*. New York: Wiley.
- Guion, R. M. (1965). *Personnel testing*. New York: McGraw-Hill.
- James, L. R., Demaree, R. G., & Mulaik, S. A. (1986). A note on validity generalization procedures. *Journal of Applied Psychology*, 71, 440-450.
- Jensen, A. R. (1984). Test validity: g versus the specificity doctrine. *Journal of Social and Biological Structures*, 7, 93-118.
- Lyman, H. B. (1978). *Test scores and what they mean* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.
- McClave, J. T., & Benson, P. G. (1982). *Statistics for business and economics*. Santa Clara, CA: Dellen.
- O'Connor, E. J., Peters, L. H., Pooyan, A., Weekley, J., Frank, B., & Erenkrantz, B. (1984). Situational constraint effects on performance, affective reactions, and turnover: A field replication and extension. *Journal of Applied Psychology*, 69, 663-672.
- Osburn, H. G., Callender, J. C., Greener, J. M., & Ashworth, S. (1983). Statistical power of tests of the situational specificity hypothesis in validity generalization studies: A cautionary note. *Journal of Applied Psychology*, 68, 115-122.
- Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology*, 65, 373-405.
- Peters, L. H., & O'Connor, E. J. (1980). Situational constraints and work outcomes: The influence of a frequently overlooked construct. *Academy of Management Review*, 5, 391-397.
- Sackett, P. R., Harris, M. M., & Orr, J. M. (1986). On seeking moderator variables in the meta-analysis of correlational data: A Monte Carlo investigation of statistical power and resistance to Type I error. *Journal of Applied Psychology*, 71, 302-310.
- Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, 62, 529-540.
- Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist*, 36, 1128-1137.
- Schmidt, F. L., & Hunter, J. E. (1984). A within setting empirical test of the situation specificity hypothesis in personnel selection. *Personnel Psychology*, 37, 317-326.
- Schmidt, F. L., Hunter, J. E., & Caplan, J. R. (1981). Validity generalization results for two groups in the petroleum industry. *Journal of Applied Psychology*, 66, 261-273.
- Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: A red herring. *Journal of Applied Psychology*, 66, 166-185.
- Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1982). Progress in validity generalization: Comments on Callender and Osburn and further developments. *Journal of Applied Psychology*, 67, 835-845.
- Schmidt, F. L., Hunter, J. E., Pearlman, K., Hirsh, H. R., Sackett, P. R., Schmitt, N., Tenopir, M. L., Kehoe, J., & Zedeck, S. (1985). Questions and answers about validity generalization and meta-analysis. *Personnel Psychology*, 38, 697-801.
- Schmidt, F. L., Hunter, J. E., Pearlman, K., & Shane, G. S. (1979). Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. *Personnel Psychology*, 32, 257-281.
- Schneider, B. (1975). Organizational climate: An essay. *Personnel Psychology*, 28, 447-479.
- Schneider, B. (1978). Person-situation selection: A review of some ability-situation interaction research. *Personnel Psychology*, 31, 281-297.
- Thorndike, R. L. (1949). *Personnel selection*. New York: Wiley.

(Appendix follows on next page)

Appendix

Table A-1
Assumed Distribution of Test and Training Criterion Reliabilities Across Studies

Reliability	Relative frequency
.90	15
.85	30
.80	25
.75	20
.70	04
.60	04
.50	02

Note. Expected value (test and training criterion reliability) = .80.

Table A-2
Assumed Distribution of Proficiency Criterion Reliabilities Across Studies

Reliability	Relative frequency
.90	03
.85	04
.80	06
.75	08
.70	10
.65	12
.60	14
.55	12
.50	10
.45	08
.40	06
.35	04
.30	03

Note. Expected value (proficiency criterion reliability) = .60.

Table A-3
Assumed Distribution of Range Restriction Effects Across Studies

Prior selection ratio	SD of test	Relative frequency
1.00	10.00	05
.70	7.01	11
.60	6.49	16
.50	6.03	18
.40	5.59	18
.30	5.15	16
.20	4.68	11
.10	4.11	05

The data in Tables A-1, A-2, and A-3 are from "Further tests of the Schmidt-Hunter Bayesian Validity Generalization Procedure" by F. L. Schmidt, J. E. Hunter, K. Pearlman, and G. S. Shane, 1979, *Personnel Psychology*, 32, pp. 260-261. Copyright 1979 by Personnel Psychology, Inc. Reprinted by permission.

Received March 21, 1986
Revision received July 7, 1986 ■

Test of the Cultural Bias Hypothesis: Some Israeli Findings

Moshe Zeidner

School of Education, University of Haifa, Mt. Carmel, Israel

The major aim of this study was to examine the cross-cultural validity of the test bias contention, with particular concern for possible sociocultural group differences in the construct and predictive validity of college entrance scholastic aptitude tests in Israel. The analyses were based on the test scores of 1,538 Israeli college student candidates of varying ethnic group membership, applying for admission to a major Israeli campus. The psychometric properties of the test battery were compared by ethnic group via a variety of internal (factor structure, reliability, etc.) and external (predictive validity, homogeneity of regression, etc.) test bias criteria. On the whole, the data provided little evidence for differential construct or predictive validity of aptitude test scores as a function of ethnic group membership, thus lending a greater deal of generality to previous research on test bias, generally negating the cultural bias hypothesis.

One of the most prevalent and serious criticisms raised by test critics against psychometric testing is that conventional aptitude tests are biased in content, procedure, and usage toward disadvantaged minority groups in the population. Therefore, it is claimed that traditional psychometric tests hold questionable validity for assessing the scholastic aptitude or predicting the scholastic performance of members of culturally different minority groups (Block & Dworkin, 1976; Cole & Bruner, 1971; Feuerstein, 1979; Ginsburg, 1972; Riessman, 1974; Sattler, 1982; Williams, 1971).

During the last 15 years or so, scholars have made substantial headway in defining the critical features of cultural bias as a scientific construct, in delineating its key facets, and in devising operational techniques for assessing the various dimensions of bias differentiated in the literature (Berk, 1982; Jensen, 1980, 1984; Samuda, 1975; Sattler, 1982; Thorndike, 1982). On the whole, recent reviews of the burgeoning test bias literature examining properly constructed standardized aptitude tests against various internal and external test bias criteria proposed in the literature, generally fail to support the cultural bias hypothesis (cf. Jensen, 1980, 1984; Reynolds, 1983).

The vast majority of studies available, using a wide variety of tests, methodologies, bias criteria, and samples, generated mainly within the American cultural scene, do not provide evidence for sociocultural group differences in construct validity

(Jensen, 1980, 1984; Reynolds, 1983; Rock, Werts, & Grandy, 1982). Likewise, most of the reviews of the literature examining sociocultural group differences in the predictive validity of aptitude test scores (e.g., Jensen, 1980; Cleary, Humphreys, Kendrid, & Wesman, 1975; Linn, 1973), provide little evidence for cultural bias in the prediction of scholastic performance. Moreover, the regression of scholastic achievement (normally assessed by first-year overall grade point average [GPA]) against scholastic aptitude test scores, is generally reported to be homogeneous across varying sociocultural groups (Jensen, 1980; Reynolds, 1983). When the regression systems do differ by culture, the difference is mainly in the intercept; that is, the minority group intercept is found to be significantly below the majority group intercept, thus, curiously resulting in the overprediction of the former group's criterion performance when the common regression line is used (cf. Jensen, 1980).

At present, however, few experts would feel confident in generalizing the results of the numerous studies carried out primarily in the American culture—focusing mainly on racial (black vs. white) group differences in internal and external test validity—beyond the American context. Indeed, one must be duly cautious when extrapolating from the available bias research data (and the theoretical models proposed to explain the available data) to culturally different contexts. The available evidence is inadequate to permit strong conclusions regarding the degree of cultural bias to be expected when applying conventional psychometric tests in cultural contexts outside the American scene. Therefore, a systematic investigation of the construct and predictive bias of psychometric tests in a cultural setting outside the American scene is indeed in order.

With respect to the Israeli scene, psychometric aptitude tests are currently used for academic selection and placement purposes by all universities and most Israeli institutions of higher education. As a rule, present testing policy in Israel espouses the use of standardized psychometric tests for assessing the scholastic aptitude of varying sociocultural groups within the Israeli population (Zeidner, 1985). Therefore, standardized scholastic aptitude tests are applied to a wide variety of subcul-

This study was made possible in part by an intramural grant from the Research Committee of the School of Education of Haifa University. Parts of the study were presented at a symposium, "Psychometric Testings: The Israeli Scene," organized and chaired by this author under the auspices of the 21st Convention of the International Association of Applied Psychology held in Jerusalem, July 12-17, 1986.

The author would like to thank two anonymous reviewers for their constructive and perspicacious insights and helpful suggestions for revision.

Correspondence concerning this article should be addressed to Moshe Zeidner, School of Education, Haifa University, Mt. Carmel, 31999, Israel.

tural groups in Israel, reflecting the pluralistic and multi-ethnic nature of Israeli society as a whole.

Since the late 1970s, however, a number of Israeli scholars (e.g., Lewis, 1979; Stahl, 1977) identified with the "cultural difference" position, have extended the cultural bias contention to the Israeli scene, claiming that commonly used ability tests may be disadvantageous to students coming from non-Western backgrounds, arguing more or less along the same lines as test critics of the cultural difference school in the American scene (cf. Mercer, 1978/1979; Samuda, 1975).

Furthermore, representatives of culturally different groups of Oriental extraction in the Israeli society, who are generally regarded as socially and educationally underprivileged when compared with their Western counterparts (Kleinberger, 1969), have repeatedly charged that scholastic aptitude tests unfairly discriminate against examinees of Oriental background. Consequently, they have called for an immediate moratorium on psychometric tests used for student selection and placement purposes (Berman, 1985). Indeed, two members of Israeli parliament of Oriental descent have recently publicly denounced college entrance aptitude tests as being culturally biased and unfair to Israeli student candidates of Asian or African backgrounds (Berman, 1985).

The antitest campaign current in Israel is also evidenced by the increasingly vehement attacks appearing in the popular press against the use of college entrance aptitude tests with lower class minority groups (e.g., Slutzki, 1985) and recent court appeals by concerned individuals attempting to restrict the use of psychometric tests for student prediction and placement purposes (E. O. Schild, personal communication, July 1985). In fact, the use of group scholastic aptitude tests within the Israeli elementary school system for purposes of student placement and selection, has been officially outlawed by the Ministry of Education as of January 1986.

Curiously, in contradistinction to the declining interest and debate surrounding the issue of *cultural bias* in the U.S.—which peaked in the late 1970s—the test bias issue in the Israeli scene has recently emerged as one of the most prominent and burning topics engaging the concern of the Israeli psychological community. In view of the fact that the bias contention has been voiced loudly with respect to college entrance aptitude tests used by all universities in the Israeli setting, it is truly peculiar that relatively little research has been devoted to the systematic scrutinization and evaluation of commonly used scholastic aptitude tests in Israel for possible cultural bias.

In the only study of predictive bias in college entrance aptitude tests published in the Israeli social science literature, Ben-Shakhar and Beller (1983) tested for the heterogeneity of the regression of GPA against aptitude test scores among students of varying ethnic background (Oriental vs. other) enrolled in their first year of studies (1974/1975) in the Faculty of Humanities and Social Sciences of the Hebrew University. The authors found no evidence of slope bias, although aptitude test scores were reported to slightly overpredict the scholastic performance of Oriental students and underpredict the performance of non-Oriental students. On the whole, the authors concluded that scholastic aptitude tests are nonbiased predictors of academic achievement among varying Jewish ethnic groups in Israel. However, in this study, test bias was examined via external criteria only. It would appear to be equally instructive to examine

Israeli scholastic aptitude tests via internal bias or construct validity criteria as well (Jensen, 1980). In addition, ethnicity was dichotomized in the Ben-Shakhar and Beller study as "Oriental" versus "others," thus lumping together students of European extraction and second-generation Israelis (about 20% of whom are of Oriental extraction) in the same category. Therefore, it would seem to be important to further test for differential predictive bias as a function of ethnic group membership, applying a more differentiated ethnic group classification scheme.

Hence, in view of the gaps in bias research, I hope to make a useful contribution to the literature by testing the cross-cultural validity of the empirical research data generated mainly within the American culture, generally negating the cultural bias hypothesis.

Specifically, the present study addresses the following questions, germane to the issue of cultural test bias in the Israeli context:

1. Do scholastic aptitude test scores measure equivalent constructs for Israeli student candidates of varying ethnic backgrounds? In particular, do scholastic aptitude tests have the same factor structure across varying ethnic subgroups?
2. Are scholastic aptitude test scores equally reliable for student candidates of varying ethnic group membership?
3. Do precollege admission tests in Israel differentially predict first-year college GPA for student candidates of varying Jewish ethnic backgrounds?

The null hypothesis, consistent with the traditional "psychometric position" (Mercer, 1978/1979; Zeidner, 1985), predicts nonsignificant ethnic group differences in the construct and predictive validity of college entrance scholastic aptitude test scores. In direct contrast, the alternative hypothesis, consistent with the cultural bias position, predicts meaningful ethnic group differences with respect to both the construct and predictive validity of aptitude tests used in the Israeli scene.

Method

Subjects

The sample consisted of 1,538 Israeli student candidates applying for admission to a major university in northern Israel for the academic year 1983/1984. They were administered scholastic aptitude tests as part of routine college entrance procedures. The analysis was based on the scholastic aptitude test scores of Jewish student candidates of identifiable ethnic group membership, who were administered the Hebrew version of the psychometric test battery during the major university testing session for the year 1983.

The sample was partitioned according to the tripartite division commonly used by researchers in Israel for classifying ethnic groups (on the basis of the father's country of origin; cf. Minkowitch, Davis, & Bashi, 1982): (a) Oriental (Asian/African), (b) European/American, and (c) Israeli. Table 1 presents the sample distribution by ethnicity and sex. Approximately 17% ($n = 262$) of the sample were second generation Israelis; about 33% ($n = 503$) were of Asian or African extraction; and about 50% ($n = 773$) were of European extraction. The sample, as a whole, was unevenly distributed by sex: The majority (61%) were women and the minority (39%), men. The group ranged in age from 17 to 68 years ($M = 24.06$, $SD = 5.61$), with examinees undifferentiated in mean age by ethnic group background. Examinee educational level was held more or less constant in our sample, with all candidates having completed at least 12 years of study.

Table 1
Sample Distribution by Ethnicity and Sex

Sex	Ethnicity			Totals
	Israeli	Oriental	European	
Male				
<i>n</i>	81	204	299	584
%	31	41	39	38
Female				
<i>n</i>	181	299	474	954
%	69	59	61	62
Totals				
<i>n</i>	262	503	773	1538
%	100	100	100	100

Note. Sex was unevenly distributed across varying ethnic groups, $\chi^2(2, N = 1,538) = 7.17, p < .05$.

Tests and Procedures

The Hebrew version of the scholastic aptitude test (SAT) battery consisted of the following six subtests, briefly described as follows:

1. General Information—40 items assessing information on a variety of topics (e.g., science, art, politics, history) particularly adapted to the Israeli cultural scene.
2. Figural Reasoning—25 items requiring examinees to identify the figure that logically continues a series of figural stimuli.
3. Mathematical Reasoning—40 problems in arithmetic, algebra, and geometry, assessing mainly numerical reasoning, problem solving, and the ability to manipulate numbers, based on the Quantitative Ability subtest of the SAT.
4. Vocabulary—30 lexical stimuli with five alternative definitions to choose from; quite similar in format to the Vocabulary subtest of the SAT.
5. Analytical Thinking—25 items requiring examinees to evaluate the validity of a variety of deductions from a given set of premises; similar in format to the analytical section of the SAT.
6. English Reading Comprehension—58 items based on five different texts in English that examinees are required to read and answer questions about.

All of the tests were of the objective multiple choice type with four to five options in each item and were administered under standardized time conditions by examiners of the university's testing unit.

Statistical Analysis

The statistical analyses of the ability test scores are based on standardized scores scaled to a mean of 50 and standard deviation of 10 for the entire sample, collapsed across ethnic groups. Ethnic group differentials in the construct validity of the scholastic aptitude test scores were assessed by a variety of techniques. To determine the stability of the factor structure across ethnic groups, scaled scores were submitted by ethnicity to principal factor analyses (R^2 in the main diagonal). An eigenvalue of 1 served as minimum criterion for factor interpretation.

To assess the comparability of factor structures across ethnic groups, coefficients of congruence were computed between corresponding factors obtained from each solution in Israeli, Oriental, and Western subgroups. The coefficient of congruence is an index of factorial similarity, taking on values from 0 to + or - 1, with a value of .90 or higher indicating equivalent factors for the groups compared (cf. Reynolds, 1982). In addition, ethnic groups were compared with respect to average intercorrelations among subtests, test communalities, percentages of total variance accounted for by factors, and comparisons of group internal con-

sistency estimates (i.e., reliability coefficients, standard errors of measurement, etc.).

Because previous research conducted by this author showed negligible sex differences with respect to the factor structure and reliability of scholastic aptitude test scores, across ethnic groups (Zeidner, in press-a) the test score data were analyzed and reported for combined groups of male and female student candidates within each ethnic group.

In order to test for possible cultural bias in the predictive validity of the scholastic aptitude test, I obtained the first-year raw cumulative GPA of all of the examinees in the applicant pool who were accepted by the university and consequently enrolled as full-time students for the academic year 1984 ($n = 696$). Unfortunately, it was unfeasible to match students from varying ethnic groups with respect to majors or courses of study, for two main reasons: (a) Students of varying ethnic background are differentially distributed across majors and courses of study, with a relatively higher proportion of European and Israeli students concentrated in the more rigorous quantitative courses of study (statistics, mathematics, computer sciences, economics, etc.) when compared with their Oriental counterparts, who are concentrated largely in the humanities (Hebrew language, literature, history, Bible studies, Near-Eastern studies), social work, and education; and (b) Israeli college students are generally required to take two different majors, therefore, matching students with respect to courses of study would have reduced the sample pool drastically. The predictive validity of psychometric tests in predicting first year GPA, as well as the regression equations of GPA against aptitude test scores, were compared for varying ethnic groups. The regression and correlational analysis for the composite test score were all based on composite scores scaled to a mean of 50 and standard deviation of 10 for the group as a whole.

Results

Profile of Ethnic Group Differences

How do Jewish student candidates of varying ethnic group membership compare with respect to their level of scholastic aptitude test performance? Table 2 presents the means and standard deviations for the composite test and individual subtests by ethnicity, along with indices (sigma units) for ethnic group size effect.

Ethnicity has a highly significant effect on total test scores, $F(2, 1535) = 58.78, p < .001$, accounting for about 7% of the composite test score variance. Statistically reliable ethnic group differences in mean test performance are also observed for each subtest included in the test battery. On the basis of squared multiple correlations between ethnicity as an independent variable and subtest score as a dependent variable, used as a descriptive statistic to describe the three group differences simultaneously, subtests rank ordered as follows with respect to ethnic group differentiation (in descending order): (a) General Information, $R^2 = .08$; (b) English Reading Comprehension, $R^2 = .08$; (c) Analytical Reasoning, $R^2 = .03$; (d) Vocabulary, $R^2 = .02$; (e) Figural Reasoning, $R^2 = .02$; and (f) Mathematical Reasoning, $R^2 = .01$.

A multivariate profile analysis (cf. Harris, 1975) testing for the parallelism of ethnic group subtest score vectors proved to be significant, $F(5, 1530) = 11.23, p < .001$, Wilks's lambda = .929. This suggests that the size of ethnic group differences in aptitude test performance is not constant across varying subtests, but varies significantly as a function of the particular subtest under consideration.

Specific post hoc comparisons were conducted for mean ethnic group differences via Dunn's procedure (Kirk, 1968). The

Table 2
Comparison of the SAT Performance of Israeli College Student Candidates by Ethnic Group Membership:
Means, Standard Deviations, and Sigma Differences

Subscale	Ethnicity						Sigma differences ^a		
	Israeli (a)		Oriental (b)		European (c)				
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>Da</i> × <i>b</i>	<i>Da</i> × <i>c</i>	<i>Dc</i> × <i>b</i>
Information	51.10	9.83	45.73	8.71	52.01	9.96	.58	−.09	.67
Figures	52.51	10.21	48.44	9.87	50.79	9.77	.42	.17	.24
Mathematics	51.37	10.22	48.32	9.85	50.63	9.84	.30	.07	.23
Vocabulary	52.49	9.87	47.99	9.16	50.39	10.29	.47	.21	.25
Analytical	52.09	9.59	47.95	10.15	51.23	9.88	.42	.09	.33
English	52.03	9.99	45.94	8.87	51.74	9.90	.65	.03	.62
Composite	52.83	9.98	46.22	9.70	51.60	9.77	.67	.12	.55

Note. *N* = 1,538. Israeli, *n* = 262; Oriental, *n* = 503; European, *n* = 773. SAT = Scholastic Aptitude Test.
^a The *D* scores designate the first group mean − the second group mean divided by the average within-group standard deviation. The three respective *D* values designate the following group differences: *Da* × *b* = Israeli–Oriental; *Da* × *c* = Israeli–European; and *Dc* × *b* = European–Oriental.

pairwise contrasts show that Israeli and European student candidates score significantly (*p* < .01) higher on average than do their Oriental counterparts, both on the composite test as well as on each of the individual subtests. In addition, the Israeli and European subgroups are not significantly differentiated with respect to total test performance, and only marginally differentiated in favor of the Israeli subgroup on the Vocabulary and Figural Reasoning scales (*p* < .05).

To determine the nature of ethnic group differences on the specific dimensions assessed by the psychometric aptitude test battery, the intercorrelation matrix among the six subtests, collapsed over ethnic groups, was submitted to a principal factor analyses (*R*² in main diagonal) with varimax rotations.

As shown in Table 3, two principal factors clearly emerge: (a) a Verbal Reasoning factor marked mainly by three subtests; General Information, Vocabulary, and English Reading Comprehension, and (b) a General Reasoning factor marked mainly by the three remaining subtests: Figural Reasoning, Mathematical Reasoning, and Analytical Thinking. On the basis of the standardized scoring coefficients derived from the factor solution, I obtained estimated factor scores for the Verbal Ability and General Reasoning factors, rescaled the factor scores to a

mean of 50 and a standard deviation of 10 for the sample as a whole, and calculated the factor score means by ethnicity (see Table 4).

Ethnicity has a significant effect on both the Verbal Ability, *F*(2, 1535) = 60.57, *p* < .001, and General Reasoning, *F*(2, 1535) = 15.95, *p* < .001, factor scores. Specific post hoc comparisons among ethnic groups via Dunn’s procedure, show that European and Israeli student candidates score significantly higher than their Oriental counterparts on both the Verbal Ability and General Reasoning factors (*p* < .01), whereas the mean scores of the former two groups are not significantly differentiated on either factor.

The two Israeli ethnic groups that are generally deemed *western* in cultural orientation—Israeli and European—appear to have a greater advantage over the Oriental subgroup on the Verbal Ability factor, as compared to the General Reasoning factor. Accordingly, there is about a 0.78 *SD* difference between Israeli and Oriental subgroup means on the verbal factor, relative to about a 0.38 *SD* difference on the General Reasoning factor. Similarly, European and Oriental subgroup means on the Verbal Ability and General Reasoning factors differ by 0.59 *SD* and 0.26 *SD*, respectively.

Table 3
Intercorrelations Among SAT Subtest Scores Collapsed Across Ethnic Groups, and Factor Structure Matrix

Subtest	Correlation matrix						Factor structure matrix		Communality (<i>h</i> ²)
	1	2	3	4	5	6	F1	F2	
1. Information	—	.207	.327	.604	.244	.541	.829	.159	.713
2. Figures		—	.498	.188	.483	.311	.095	.733	.646
3. Mathematics			—	.218	.429	.345	.220	.614	.426
4. Vocabulary				—	.274	.431	.692	.150	.501
5. Analytical					—	.433	.219	.638	.455
6. English						—	.541	.392	.446
% total variance							26	25	

Note. All correlations are significant at the .05 level. The two factors that emerged from a principal factor analysis of the sample intercorrelation matrix were labeled Verbal Ability (F1) and General Reasoning (F2). SAT = Scholastic Aptitude Test.

Table 4
Factor Score Means and Standard Deviations by Ethnicity, and Ethnic Group Effect Size Estimates

Factor	Ethnicity						Effect size ^a		
	Israeli (a)		Oriental (b)		European (c)		$Da \times b$	$Da \times c$	$Dc \times b$
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>			
Verbal	53.33	9.83	45.97	8.94	51.60	10.03	.78	.17	.59
Reasoning	52.19	10.24	48.35	9.97	50.93	9.61	.38	.13	.26

Note. $N = 1,538$. Israeli, $n = 262$; Oriental, $n = 503$; European, $n = 773$. Estimated factor scores were scaled to mean of 50 and a standard deviation of 10 for the entire sample collapsed across ethnic groups.

^a The D scores designate the first group mean minus the second group mean divided by the average within-group standard deviation. The three respective D values designate the following group differences: $Da \times b$ = Israeli–Oriental; $Da \times c$ = Israeli–European; and $Dc \times b$ = European–Oriental.

Ethnic groups tend to evidence consistent rank order with respect to level of test performance across most aptitude scales and factor scores, with Israeli student candidates generally coming out on top, followed closely by European student candidates, and trailed by their Oriental counterparts.

Ethnic Group Differences in Construct Validity

To make fair and consistent interpretations of aptitude test scores across varying ethnic groups, the scales under consideration must be shown to measure the same constructs with the same degree of accuracy.

Do psychometric aptitude test scores differ in factor structure for student candidates of varying ethnic background? In order to address this question, I subjected the six subtests composing the SAT battery to a principal factor analysis with varimax rotations by ethnicity. The communalities appearing in Table 3, derived from the principal factor analysis for the entire sample, were used as estimates of communality and were placed on the main diagonal of each subgroup intercorrelation matrix. Table 5 shows the intercorrelations among the six ability subtests and the subtest loadings on the two principal factors for student candidates of Israeli, Oriental, and European backgrounds.

The central tendencies and dispersions of the correlations among the subtests are highly similar for Israeli ($M = 0.38$, $SD = 0.13$), European ($M = 0.34$, $SD = 0.13$), and Oriental ($M = 0.38$, $SD = 0.11$) subgroups. Furthermore, product-moment correlations were calculated between the 15 strung out analogous correlations appearing in the intercorrelation matrices for various combinations of the three ethnic subgroups. The observed correlations ranged between .87 and .95, as follows: (a) Israel \times Oriental, $r = .95$, $p < .001$, (b) Israel \times European, $r = .87$, $p < .001$, and (c) Oriental \times European, $r = .90$, $p < .001$. In addition, an approximate test suggested by Jensen (1980) for the equivalence of correlation matrices across samples from different populations, shows that the correlation matrices are equivalent for all group comparisons (i.e., Israel \times Oriental, Israel \times European, Oriental \times European).

Thus, in view of the high degree of similarity between the correlation matrices of the various ethnic groups, one would also expect to find a high degree of similarity between the factor structures of the three ethnic groups. An inspection of Table 5 shows that indeed this is the case. Accordingly, principal fac-

tor analyses of the intercorrelation matrices by ethnic group yield the same two nontrivial components, Verbal Ability and General Reasoning, in each of the three groups. The two common principal components together explain about 50% of the total variance in each group.

The degree of factor similarity for varying ethnic subgroups was gauged via coefficients of congruence, calculated between all possible pairwise comparisons among ethnic groups for corresponding components. The observed congruence coefficients for the Verbal Ability and General Reasoning factors, ranging from .98 to .99, varied little indeed, and are of such magnitude as to represent virtual identity of both components across ethnic groups.

Furthermore, for a test to be considered culturally nonbiased according to construct validity criteria, equivalency in the internal consistency and accuracy of the test scores across groups would be required (Reynolds, 1982). Table 6 presents the subtest and composite test KR-20 reliability coefficients and respective standard errors of measurement calculated separately by ethnicity. A casual examination of Table 6 shows a high degree of similarity in both the composite test and subtest reliability coefficients and standard errors of measurement for student candidates of varying ethnic group background. Accordingly, the maximum difference between ethnic groups in subtest reliability coefficients for any given subtest ranges from about .02 to .12 reliability coefficient points; the ethnic group differences in standard errors of measurement units ranges between .08 and .32 points for any given subtest. Therefore, on the whole, the accuracy of test scores assessed by standard errors of measurement, appears to be highly comparable across ethnic groups, with the observed differences in reliability indices judged to be of negligible practical consequences. The low reliabilities for the Vocabulary test scores in each of the ethnic subgroups is unusual and not readily explicable.

Ethnic Group Differences in Predictive Validity

Although a given scholastic aptitude test or test battery may evidence homogeneous construct validity across varying socio-cultural groups, the test may still be found to be a culturally biased predictor of scholastic achievement for particular ethnic subgroups (Jensen, 1980). Therefore, in order to test for ethnic group differences in the predictive validity of psychometric ap-

Table 5

Intercorrelation Matrix of Scholastic Aptitude Subscales and Factor Structure Matrix by Ethnicity

Subtest	Correlation matrix						Factor loadings		Communality (h^2)
	1	2	3	4	5	6	F1	F2	
Israeli student candidates ($n = 262$)									
1. Information	—	.22	.30	.62	.26	.46	.83	.16	.70
2. Figures		—	.52	.18	.53	.33	.08	.79	.63
3. Mathematics			—	.28	.51	.32	.23	.63	.44
4. Vocabulary				—	.34	.41	.70	.18	.52
5. Analytic					—	.45	.27	.66	.50
6. English						—	.51	.39	.41
% total variance							26	28	
Oriental student candidates ($n = 503$)									
1. Information	—	.28	.39	.56	.25	.48	.81	.20	.70
2. Figures		—	.54	.27	.51	.37	.16	.78	.64
3. Mathematics			—	.28	.40	.39	.26	.60	.43
4. Vocabulary				—	.24	.42	.67	.17	.48
5. Analytic					—	.42	.19	.64	.44
6. English						—	.51	.43	.44
% total variance							25	27	
European student candidates ($n = 773$)									
1. Information	—	.18	.32	.63	.22	.47	.79	.14	.70
2. Figures		—	.46	.17	.41	.25	.08	.77	.59
3. Mathematics			—	.21	.40	.31	.20	.59	.40
4. Vocabulary				—	.28	.43	.76	.15	.52
5. Analytic					—	.43	.23	.59	.41
6. English						—	.51	.35	.42
% total variance							24	27	

Note. All correlations are significant at the .05 alpha level.

* Factors 1 and 2 designate the Verbal Ability and General Reasoning factors, respectively, derived from a principal factor analysis followed by varimax rotations.

aptitude tests, I obtained the first-year cumulative GPA of 696 students (about 45% of the original applicant pool), constituting all examinees in the student applicant pool who were enrolled as full-time university students during the academic year 1984.

Table 7 presents the summary statistics for the test variables and GPA, by ethnicity in the student subsample. A comparison of aptitude test scores for the original examinee pool (see Table 2) and the restricted student pool (see Table 7) shows with few exceptions, highly similar mean subtest profiles by ethnic group membership. Also, ethnic groups rank order identically with respect to GPA and composite test scores. Accordingly, Israeli students show the highest mean GPA, followed by European and Oriental students.

Pearsonian product-moment correlations between first-year GPA as criterion and composite test scores as predictor, were uniformly significant but modest within Israeli ($r = .22$), Oriental ($r = .27$), and European subgroups ($r = .20$). The correlations were not statistically heterogeneous.

Significant multiple correlation predictive validities are found for all three ethnic subgroups: Israeli ($R = .35$), Oriental ($R = .42$), and European ($R = .28$). A test for the statistical heterogeneity of the multiple correlation validity coefficients by ethnicity, following R to z transformations (cf. Reynolds, 1982), shows that they are not significantly differentiated, $\chi^2(1, N = 696) = 3.25, p > .05$. In addition, based on a test of significance

suggested by Reynolds (1982), the standard errors of estimate obtained separately by ethnicity, are not significantly different for Israeli (9.30), Oriental (10.55), and European (10.08) subgroups.

Table 6

Reliability Coefficients (KR-20) and Standard Errors of Measurement for Subscales by Ethnic Group

Subscale	Ethnicity					
	Israeli		Oriental		European	
	r_{xx}	SEM	r_{xx}	SEM	r_{xx}	SEM
Information	.803	4.36	.814	3.76	.819	4.23
Figures	.762	4.98	.708	5.33	.709	5.27
Mathematics	.831	4.20	.813	4.26	.789	4.52
Vocabulary	.638	5.93	.563	6.06	.680	5.82
Analytical	.809	4.19	.818	4.33	.802	4.40
English	.897	3.21	.878	3.10	.900	3.13
Composite	.928	2.68	.950	2.17	.920	2.76

Note. The standard errors of measurement (SEM) computed were based on KR-20 reliability coefficients (r_{xx}). The composite score reliabilities were calculated using Mosier's (1943) formula.

Table 7
Scaled Aptitude Test Scores and Grade Point Average Means, Standard Deviations, and Effect Size Estimates for Student Subsample by Ethnic Group Membership

Subscale	Ethnicity						Sigma differences ^a		
	Israeli (a)		Oriental (b)		European (c)				
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>D</i> a × b	<i>D</i> a × c	<i>D</i> c × b
Information	52.57	9.21	46.97	9.52	53.19	10.08	.60	−.06	.63
Figures	52.18	9.74	50.84	9.38	50.92	9.71	.14	.13	.08
Mathematics	51.72	9.22	48.95	10.22	51.45	9.99	.28	.03	.25
Vocabulary	53.81	9.04	49.61	9.81	51.81	10.11	.44	.21	.22
Analytical	52.81	8.87	49.77	9.81	51.31	9.86	.33	.16	.16
English	51.89	8.92	47.01	8.90	53.19	10.31	.65	−.14	.62
Composite	53.72	8.16	48.53	9.28	52.96	9.67	.60	.09	.47
GPA	77.01	9.93	71.32	11.80	73.67	11.53	.52	.31	.20

Note. GPA = grade point average. *N* = 696. Israeli, *n* = 131; Oriental, *n* = 228; European, *n* = 337. Raw score GPAs are presented in the table.
^a The *D* scores designate the first group mean minus the second group mean divided by the average within-group standard deviation.

Furthermore, I calculated the differences between observed and predicted mean GPAs (observed–predicted), based on the group multiple regression equation, and expressed the differences in terms of GPA standard deviation units by ethnic group (Jensen, 1980). For Israeli student candidates, the observed scores were about +.28 *SD* above what would be predicted from the common regression line. For Oriental student candidates, in direct contrast, the mean difference between observed and predicted GPA was −.14 *SD* units below what would be predicted on the basis of the common regression line. For European student candidates, the mean difference between observed and predicted GPA was nil. On the whole, the magnitude of the differences between observed and predicted mean GPA was small within each of the ethnic groups. However, the intercept of the Oriental group (49.42) was meaningfully below that of both Israeli (61.08) and European (57.07) students.

Note that the multiple correlation predictive validities reported already by ethnic group, may be inflated by capitalizing on chance (i.e., sampling error). Therefore, they were corrected for the expected shrinkage that would occur if the regression weights were applied to a new sample, using the formulas provided by Cattin (1980). The estimated population cross-validated multiple correlation coefficients for Israeli, Oriental, and European subgroups were .23, .38, .22, respectively. It is noteworthy that the cross-validated multiple correlation validity coefficients obtained are much more similar to the validities given for composite scores, as one would expect.

Table 8 presents the regression constants for GPA, regressed on scaled composite aptitude test scores as well as on each of the six individual aptitude subtests by ethnic groups. Figure 1 presents the regression plots of GPA as a function of composite aptitude test scores for the three groups.

A test for ethnic group differences in the regression of GPA on composite aptitude test scores, given the common regression line, was conducted via analysis of covariance (ANCOVA) procedure, as suggested by Freund and Littell (1981). Accordingly, data was pooled across ethnic groups to allow for consideration of a single regression model in predicting GPA, consisting of the following three terms: composite test scores, ethnicity, and the

Ethnicity × Composite Test Scores interaction. The equal intercept hypothesis was examined by testing the significance of the ethnicity term in the model, whereas the equal slope hypothesis was examined by testing the significance of the interaction term. The ANCOVA reveals that ethnicity does not significantly interact with aptitude test scores in predicting GPA, $F(2, 683) = 1.74, p = .18, ns$. Thus, no significant differences in the regression relationship by ethnic group are indicated. However, ethnicity bears a significant effect on GPA, $F(2, 683) = 8.41, p < .001$, indicating the presence of a slight degree of intercept bias (see Figure 1).

The intercept differences in standard deviation units for the total test score regressions for the three groups were highly comparable to those obtained when using the subtest regression equations already detailed. Accordingly, for Israeli student candidates, the observed GAP scores were about +0.26 *SDs* above what would be predicted from the common regression line. For Oriental student candidates, in direct contrast, the mean difference between observed and predicted GPA was −0.15 *SD*. The degree of underprediction was practically nil for European students, amounting to about +0.07 *SD*.

As shown in Table 8, the intercept for the Oriental subgroup ($a = 54.67$) appears to be meaningfully below that of the Israeli ($a = 62.61$) and European subgroups ($a = 61.06$). In order to determine the extent to which intercept differences may be due to unreliability in composite test score, the true score intercepts were computed for each group (cf. Hunter & Schmidt, 1976; Jensen, 1980). The true score intercepts for Israeli, ($Ct = 61.53$) Oriental ($Ct = 53.95$), and European ($Ct = 59.95$) subgroups suggest that the intercept differences are real and do not appear to be due to unreliability in total test score.

Figure 2 presents the regression lines for GPA, regressed on each of the six tests by ethnic group (see Table 8 for the regression constants). In addition, separate tests conducted for ethnic group differences in the regression of GPA against each of the six subtests appearing on the test battery show significant ethnic group differences in slopes for the English subtest only, $F(2, 683) = 6.53, p < .002$. Thus, the slope of GPA regressed on English Aptitude for the European group ($b = .15$) is meaning-

Table 8
Bivariate Regression Constants of First-Year Cumulative GPA
Regressed on Composite Test Score and Individual
SAT Subtest Scores by Ethnicity

Subscale	Ethnicity					
	Israeli		Oriental		European	
	<i>a</i>	<i>b</i>	<i>a</i>	<i>b</i>	<i>a</i>	<i>b</i>
Information	75.96	.02	58.64	.27	65.76	.15
Figures	72.84	.08	71.32	.00	71.86	.04
Mathematics	77.01	.00	68.49	.06	72.42	.02
Vocabulary	71.09	.11	59.38	.24	61.85	.23
Analytical	67.55	.18	59.95	.23	64.67	.18
English	60.84	.31	48.26	.49	65.97	.15
Composite	62.61	.27	54.67	.34	61.06	.24

Note. GPA = grade point average. SAT = Scholastic Aptitude Test. *a* and *b* designate the intercept and slope constants, respectively.

fully and inexplicably below that of the Israeli ($b = .31$) and particularly the Oriental subgroup ($b = .49$). (See Table 8 and Figure 2, respectively.) Furthermore, significant intercept differences were found on all six subtests: (a) Information, $F(2, 683) = 7.33, p < .001$; (b) Figural Reasoning, $F(2, 683) = 10.54, p < .001$; (c) Mathematical Ability, $F(2, 683) = 10.38, p < .001$; (d) Vocabulary, $F(2, 683) = 7.59, p < .001$; (e) Analytical Thinking, $F(2, 683) = 8.82, p < .001$; and (f) English, $F(2, 683) = 5.98, p < .01$.

In addition, it is interesting to note that when ethnic group membership is entered into the regression equation following composite aptitude test scores, it contributes significantly to the prediction of GPA, raising the bivariate validity coefficient from .21 to .25 and contributing about 2% to the prediction of GPA, above and beyond aptitude test scores.

Past research in Israel has shown significant sex group differences in total aptitude test performance, with men scoring about .27 sigma units above women (Zeidner, in press-a). In the present research, the product-moment correlations between GPA and composite aptitude test scores by sex, collapsed across ethnic groups, were comparable and statistically nondiscernible for male ($r = .24, p < .001$) and female ($r = .30, p < .001$) students. A test for the heterogeneity of slopes by sex group proved to be nonsignificant. However, the ANCOVA showed the intercept of female students ($n = 438; a = 61.08$) to be significantly higher than those of male students ($n = 258; a = 54.94$), suggesting that the GPA of female Jewish students is underpredicted by the common regression line. Thus, female GPA is observed to be about 0.26 *SD* above what would be predicted from the common regression line, whereas male GPA is about 0.37 *SD* below what would be predicted from the common regression line.

The validity coefficients and intercept differences were calculated separately by ethnicity and sex: (a) Israeli men ($n = 45$), $r = .16, d = -.29$; (b) Israeli women ($n = 86$), $r = .27, d = +.82$; (c) Oriental men ($n = 85$), $r = .27, d = -.46$; (d) Oriental women ($n = 121$), $r = .35, d = .06$; (e) European men ($n = 102$), $r = .21, d = -.33$; and (f) European women ($n = 205$), $r = .20, d = +.27$. Thus, although the performance of men within each

ethnic group appear to be overpredicted, the performance of women is meaningfully underpredicted, mainly in the Israeli subgroup.

In addition, the intercept differences for ethnic groups were calculated separately in male and female subgroups. In the male student group, the intercept differences were $+.08, -.06$, and $+.02$ for Israeli, Oriental, and European examinees, respectively; in the female subgroup, they were $+.42, -.13$, and $-.02$ for Israeli, Oriental, and European examinees, respectively. Thus, once again, female Israeli students appear to be particularly underpredicted by the common (within sex group) regression line.

Discussion

The major goal of the present study was to examine the cross-cultural validity of the bias contention when put to an empirical test in the Israeli scene. To that end, I explored the possible ethnic group differences in the construct and predictive validity of scholastic aptitude tests routinely used for academic selection purposes in the Israeli scene.

The present investigation provided substantial evidence for ethnic group equivalence in the factor structure of college entrance scholastic aptitude subtests, as well as for comparable test score reliability across ethnic groups. Thus, this study concurs with the bulk of data generated in the American scene demonstrating similar construct validity across ethnic groups. As Reynolds (1983) has concluded on the basis of a survey of the bias research literature, differential construct validity has not been found and is not likely to be found to be a major phenomenon in properly constructed psychometric tests. Therefore, from the point of view of construct validity, this study suggests that college entrance aptitude tests applied in the Israeli scene, measure the same constructs and function with the same degree of accuracy across varying ethnic subgroups. Thus, aptitude tests do not appear to meaningfully fuzz or distort true ethnic group differences in aptitude test performance, with the observed ethnic group differences most likely real and not merely artifactual.

The data support the notion that ethnic group differences in mean ability test performance in Israel are largely accounted for by the verbal dimension of the test battery, which is heavily

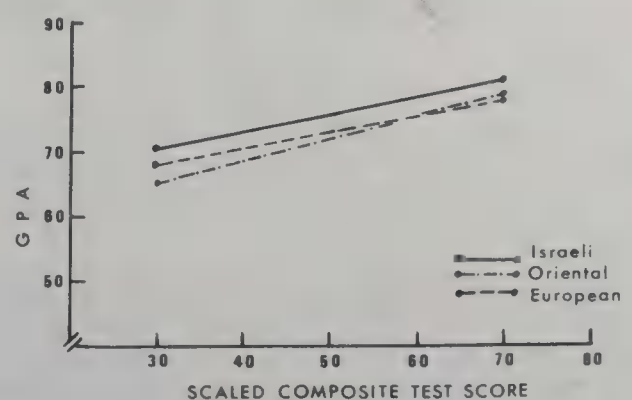


Figure 1. Regression of first-year cumulative grade point average (GPA) on composite aptitude test in Israeli, Oriental, and European student subgroups.

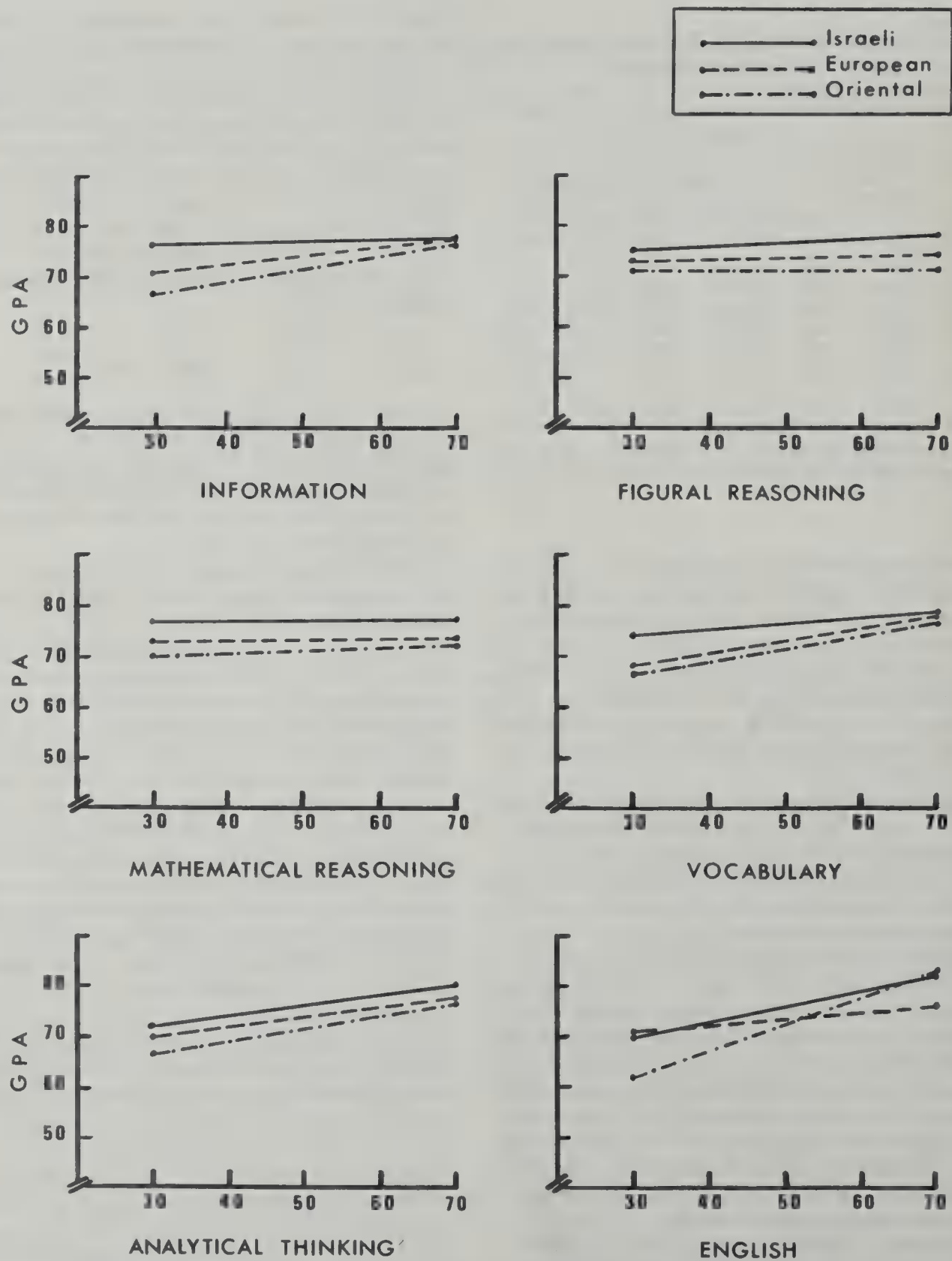


Figure 2. Regression of first-year cumulative grade point average (GPA) on individual aptitude test in Israeli, Oriental, and European student subgroups.

influenced by the examinee's past learning experiences. However, verbal deficit is not the locus of the problem for all minorities. For example, blacks in the United States differ from whites as much on the Wechsler performance test score as they do on the Verbal test score (Jensen, 1980).

Furthermore, our study provides little evidence for ethnic group differences in the predictive validity of aptitude tests. On the whole, the results of this study are consistent with much of the

research reported in the literature, showing little evidence for slope bias but some evidence for intercept bias when scholastic aptitude test scores are used in predicting GPA (Jensen, 1980). Thus, similar to previous studies reported in the literature, the minority group intercept was observed to be slightly below that of the majority group, which, curiously, would result in the overprediction of the minority group's criterion performance when the common or majority group regression line is used.

The data presented are perfectly consistent with those presented by Ben-Shakhar and Beller (1983) within the Israeli scene, pointing to an overprediction of first year cumulative GPAs for Oriental students and an underprediction of GPAs for non-Oriental students.

This study illustrates how ubiquitous the phenomenon of overprediction of minority performance really is, and underlines the importance for psychologists engaged in student selection and placement to be fully cognizant of this widespread phenomenon. As Jensen (1980) has pointed out, when the mean of the majority group on the predictor is above that of the minority group, systematic errors of prediction may occur purely because of the unreliability of the test, given a common regression line. In fact, when overprediction is found for a fallible measure, reducing the amount of measurement error in the predictor would concomitantly decrease the amount of overprediction.

One of the most surprising findings in this study is the difference in mean college grades of the Israeli and European groups compared to their near equality on the tests. In view of the disproportionate sex ratios in the two groups and the meaningful underprediction of Israeli female GPA., it appears that the mean differences may be due to the extra number of women in the Israeli group who perform meaningfully above what would be predicted from the common regression line. This observed sizable intercept difference within the female Israeli group may be due to a number of factors, including the concentration of Israeli women in certain courses in the humanities that have particularly easy grading standards; the particularly high motivational levels of Israeli women and their language fluency, which facilitates success in academic work; and the relatively small GPA standard deviation in that group (6.81), as compared to the larger standard deviations (ranging from 8.90 to 11.93) in the other subgroups.

It is important to add that the relatively low validities obtained in this study may be due to major differences in grading standards in different majors. Validity coefficients calculated within majors (or within dichotomized categories of predominant majors, namely, "hard majors" vs. "soft majors") would be expected to be higher than have been shown thus far and would probably give a truer picture of the validity and value of the test for the group as a whole.

When one compares the subgroup means and standard deviations in the applicant group with those in the group that entered the university for full-time study, one is struck by the high degree of similarity in the mean and standard deviation subtest profiles. Thus, the question may be asked, What purpose is actually served by the entrance tests?

First, although aptitude tests in concert with matriculation certificate grades are used as primary criteria for student selection purposes by all Israeli universities, most university departments set departmental cutoff scores for student acceptance so that, in effect, the tests are used no less in a classification than in a selection capacity.

Second, note that a large percentage of applicants scoring in the middle range of the test distribution, who would be eligible for study in the university, choose not to register at the university under consideration, but find other alternatives for pursuing their future education (teacher training, art school, technical studies, other universities in Israel). In addition, many students

with more than middle-level ability are attracted to medical and engineering schools in the nearby vicinity and attend these rather than the university.

Third, it should be recalled that the present study focuses on cultural bias in aptitude tests for various ethnic groups in the Jewish population. Given the relatively small magnitude of group differences in aptitude test performance in the Jewish population, a relatively small percentage fail to meet the university cutoff score for acceptance. However, a larger percentage of non-Jewish minority group student candidates (e.g., Druze, Arab) in the university under consideration, who score about one standard deviation below Jewish students on composite test scores (Zeidner, *in press-b*), would have failed to meet the cutoff point for university selection; this would be manifested among non-Jewish minorities, in more marked differences in mean and standard deviation profiles for the candidate pool compared to the selected student group.

In the present study, bias was conceived and operationally defined as systematic error in the construct or predictive validity of psychological tests associated with sociocultural group membership. This definition is perfectly consistent with the predominant conceptualization of bias in the psychometric literature (Jensen, 1980; Reynolds, 1983; Thorndike, 1982). However, it should be pointed out that there are alternative claims to the definition of bias (e.g., unfamiliar test content, significant mean differences, overinterpretation of test results, inappropriate criterion, inappropriate national norms, aversive test atmosphere, inequitable social consequences of tests), with well over a dozen definitions and dimensions of bias delineated in the psychological literature (cf. Jensen, 1980; Sattler, 1982). Thus, research based on alternative definitions of test bias or different research paradigms may yield results and conclusions different from those reported herein.

In view of the meaningful overlap between ethnic group and social class background in Israeli society as a whole (Kleinberger, 1969), it is highly probable that the ethnic groups compared in this study, vary in terms of family socioeconomic status (SES), with a larger percentage of Oriental examinees predicted to come from lower class family background, relative to their European or Israeli counterparts. Unfortunately, because no measure of SES was on record, it was impossible to determine the magnitude of social class variance among the ethnic groups included in our sample. Nevertheless, the question addressed in this study is whether measured ability can account for the differences among the groups in academic performance. This is essentially what was found: Mean academic performance for the Oriental group was at least as low (and actually slightly lower) as predicted by test scores. The finding reported in this study of essentially equal regression lines across groups differing on both SES and ethnicity means that neither SES nor ethnicity is relevant to the relation between ability and performance. That is, neither affects performance except through its impact on ability.

In sum, keeping in mind the caveats cited, the cultural bias hypothesis—contending that standardized aptitude tests are systematically biased against minority groups—was once again disconfirmed, but now in the Israeli cultural context, thereby generalizing the results beyond the American scene. On the whole, the data provide little evidence to suggest that psycho-

metric aptitude tests are biased or unfair toward any of the ethnic groups included in this study, thus providing increased confidence in the appropriateness of applying the tests for decision purposes among student candidates of both minority and majority ethnic group background in Israel. Therefore, current university admissions policy in Israel advocating the use of aptitude tests for student candidates of Oriental and Western groups alike, need not be altered in any major way.

Although cultural bias may not be a key determinant of observed group differences in test performance, it nevertheless needs to be ruled out in our pursuit of understanding the origins of test performance deficiencies among varying ethnic groups. The present study provides, in part, that ruling out with respect to the Israeli academic scene and also provides an answer to those who have contended that cultural bias in scholastic aptitude testing largely explains observed sociocultural mean differences in test performance. Therefore, in view of the results, it may now be wise to direct our attention to more critical variables, in pursuit of understanding the origins of ethnic disparities in aptitude test performance within the Israeli scene. Indeed, psychologists engaged in applied settings have often been guilty of an exaggerated concern with the potential bias in aptitude tests. It may be argued instead that more attention should be devoted to increasing criterion performance, which in effect may be the main locus of the minority group problem.

References

- Ben-Shakhar, G., & Beller, M. (1983). On the cultural fairness of psychological tests. *Megamot Behavioral Sciences Quarterly*, 28, 42-56.
- Berk, R. A. (1982). (Ed.). *Handbook of methods for detecting test bias*. Baltimore, MD: Johns Hopkins Press.
- Berman, I. (1985, May). *Is the psychometric examination biased?* Paper presented at the meeting of the Academic Committee for Research on Language Testing, Kiryat Anavim, Israel.
- Block, N. J., & Dworkin, G. (1976). IQ, heritability and inequality. In N. J. Block, & G. Dworkin, *The IQ controversy* (pp. 410-540). New York: Pantheon Books.
- Cattin, P. (1980). Estimation of the predictive power of a regression model. *Journal of Applied Psychology*, 65, 407-414.
- Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. (1975). Educational uses of tests with disadvantaged students. *American Psychologist*, 30, 15-41.
- Cole, M., & Bruner, J. S. (1971). Cultural differences and inferences about psychological processes. *American Psychologist*, 26, 867-876.
- Feuerstein, R. (1979). *The dynamic assesment of retarded performers*. Baltimore, MD: University Park Press.
- Freund, R. J., & Littell, R. C. (1981). *SAS for linear models: A guide to the ANOVA and GLM procedures*. Cary, NC: SAS Institute.
- Ginsburg, H. (1972). *The myth of the deprived child*. Englewood Cliffs, NJ: Prentice-Hall.
- Harris, R. J. (1975). *A primer of multivariate statistics*. New York: Academic Press.
- Hunter, J. E., & Schmidt, F. L. (1976). Critical analyses of the statistical and ethical implications of various definitions of *test bias*. *Psychological Bulletin*, 83, 1053-1071.
- Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.
- Jensen, A. R. (1984). Test bias: Concepts and criticisms. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing* (pp. 507-586). New York: Plenum Press.
- Kirk, R. E. (1968). *Experimental design: Procedures for the behavioral sciences*. Belmont, CA: Wadsworth.
- Kleinberger, A. F. (1969). *Society, schools and progress in Israel*. London: Pergamon Press.
- Lewis, A. (1979). *Power, poverty and education*. Ramat-Gan, Israel: Turtledove Publishers.
- Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research*, 43, 139-161.
- Mercer, J. R. (1978/1979). Test validity, bias and fairness: An analysis from the perspective of the sociology of knowledge. *Interchange*, 9, 1-16.
- Minkowitch, A., Davis, D., & Bashi, Y. (1982). *Success and failure in Israeli elementary education*. New Brunswick, NJ: Transaction Books.
- Mosier, C. I. (1943). On the reliability of a weighted composite. *Psychometrika*, 8, 161-168.
- Reynolds, C. R. (1982). Methods for detecting construct and predictive bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 199-227). London: Johns Hopkins Press.
- Reynolds, C. R. (1983). Test bias: In God we trust, all others must have data. *Journal of Special Education*, 17, 241-260.
- Riessman, F. (1974). The hidden IQ. In A. Gartner, C. Greer, & F. Riessman (Eds.), *The new assault on equality: IQ and social stratification* (pp. 206-223). New York: Harper & Row.
- Rock, D. A., Wertz, C., & Grandy, J. (1982). *Construct validity of the GRE aptitude test across populations—an empirical confirmatory study* (Research Report No. 81-57). Princeton, NJ: Educational Testing Service.
- Samuda, R. J. (1975). *Psychological testing of American minorities*. New York: Harper & Row.
- Sattler, J. M. (1982). *Assessment of children's intelligence and special abilities* (2nd ed.). Boston: Allyn & Bacon.
- Slutzki, S. (1985, October). Examinations as obstacles. *Al Hamishmar*, pp. 6-7.
- Stahl, A. (1977). *The language and thought of culturally disadvantaged students in Israel*. Tel Aviv, Israel: Otsar HaMoreh.
- Thorndike, R. L. (1982). *Applied psychometrics*. Boston: Houghton Mifflin.
- Williams, R. L. (1971). Abuses and misuses in testing black children. *Counseling Psychology*, 2, 62-67.
- Zeidner, M. (1985). A cross-cultural test of the situational bias hypothesis—The Israeli scene. *Evaluation and Program Planning*, 8, 267-276.
- Zeidner, M. (in press-a). Sex differences in scholastic aptitude: The Israeli scene. *Personality and Individual Differences*.
- Zeidner, M. (in press-b). Validity of college admission indices for Jews and Arabs in Israel. *Personality and Individual Differences*.

Received June 2, 1986 ■

Comparison of Several Procedures for Generating J-Coefficients

John W. Hamilton
DOW Chemical USA, Midland, Michigan

Terry L. Dickinson
Old Dominion University

J-coefficients were examined as estimators of criterion-related validities for the occupation of machinist in a medium-sized international manufacturer of precision automotive accessories. Test and job performance components of the J-coefficient were estimated from supervisory, incumbent, co-worker, policy capturing, importance, and test expert ratings. The results indicated that combinations of estimates from several sources generated J-coefficients that reproduced the pattern and mean values of the validities. Supervisory and incumbent sources were superior to the test expert source for estimating test components, whereas co-worker, policy capturing, and importance sources were equally effective for estimating job performance components. The implications of the J-coefficient are discussed for personnel selection.

The J-coefficient has been recommended for estimating criterion-related validities when (a) sample sizes are inadequate to conduct criterion-related studies, (b) jobs are new and without incumbents, (c) jobs change rapidly in their content, and (d) adequate measures of performance are unavailable (Balma, 1959; Primoff, 1964; Wherry, 1955). As suggested by Guion (1976, p. 821), the J-coefficient holds promise for providing a "scientific basis for the art of decision" in selection research.

Statistical Definitions

Fundamental to the J-coefficient is the breakdown of job performance into its constituent elements. These elements may be specified with a job analysis as behaviors, traits, abilities, or skills, but they must be conceivable as determinants of job performance (Wherry, 1955). For any given job it is usually assumed that not all of the determinants of job performance can be identified (Dickinson, 1976). Some job elements may not be specified by the job analysis, and many of the situational variables that modify job performance remain unknown. However, the J-coefficient can correct for the underidentification of job elements and their moderators. More specifically, the J-coefficient is defined by the following two equivalent formulas:

$$J_{XY} = \frac{B_X^* R_{YE}}{(R_{XE}' R_{EE}^{-1} R_{XE})^{1/2} (R_{YE}' R_{EE}^{-1} R_{YE})^{1/2}}, \quad (1)$$

or

$$J_{XY} = \frac{B_Y^* R_{XE}}{(R_{XE}' R_{EE}^{-1} R_{XE})^{1/2} (R_{YE}' R_{EE}^{-1} R_{YE})^{1/2}}, \quad (2)$$

where

B_X^* = row matrix of standard score regression coefficients from the regression of test performance on the job elements;

B_Y^* = row matrix of standard score regression coefficients from the regression of job performance on the job elements;

R_{XE} = column matrix of correlations between test performance and the elements;

R_{YE} = column matrix of correlations between job performance and the job elements; and

R_{EE}^{-1} = inverse matrix of correlations among job elements.

The purpose of this study was not to report on the derivation or extensions of these formulas, because this has been accomplished elsewhere (see Dickinson, 1976; Primoff, 1955; Urry, 1978; Wherry, 1955). Instead, we focused our attention on procedures for estimating the component relations that define the J-coefficient,

B_X^* , B_Y^* , R_{XE} , R_{YE} , and R_{EE}^{-1} .

Estimating the Components

The promise of the J-coefficient for personnel selection is based on the fact that a validity coefficient can be defined by the component relations (a) between job elements and test performance, (b) between job elements and job performance, and (c) among job elements. Although these component relations could be measured with criterion-related validation designs, such designs typically correlate test and job performance directly and do not measure component relations. If the designs are not feasible, because test and job performance cannot be measured directly, then the J-coefficient is an alternative. Judgmental or empirical estimates of component relations can be used to generate the J-coefficient as an estimator of the validity coefficient.

When using judgments to obtain estimates of component relations for generating J-coefficients, both the source and the procedure must be considered. The logical sources of judg-

The authors would like to thank three anonymous reviewers for their helpful comments.

Correspondence concerning this article should be addressed to John W. Hamilton, Psychology and Training Resources, DOW Chemical USA, 47 Building, Midland, Michigan 48640.

mental estimates are experts who are familiar with the job or test. Job incumbents or supervisors would be expected to be more familiar with the elements constituting the job and how they are related to performance, whereas test experts would be expected to provide more knowledgeable judgments about test performance.

Two judgmental procedures have been used to estimate the relations between job elements and job performance. Primoff (1955, 1959) used job experts to rate the importance of 62 job elements to job performance. These job elements were developed by the U.S. Employment Service for use as a worker characteristics checklist. The checklist was later modified by the U.S. Civil Service Commission, and it was meant to be applicable to a broad range of jobs. The R_{YE} component for generating a J-coefficient is obtained by an arithmetic averaging of the ratings of the importance of the elements to performing the job. Policy capturing is a second procedure that has been used for estimating R_{YE} , as well as B^* (Dickinson & Wijting, 1976; Mullins & Usdin, 1970). Job experts rate the job performance of hypothetical incumbents, who are described by their profiles of job element scores. The ratings are regressed on the job element scores to capture each rater's policy, B^* , and then these policies are clustered to obtain composite estimates of B^* and R_{YE} for generating a J-coefficient.

In his early work, Primoff (1955, 1959) estimated B^* empirically rather than judgmentally. This was done through criterion-related validation studies for a battery of tests that was developed by the Civil Service Commission to be used with the 62 job elements. A table of B^* components was developed and continually updated for a variety of occupations for each test and the 62 job elements. The B^* component for generating a J-coefficient is simply chosen from the table for the same or a similar occupation. In contrast, experts could be used to estimate the relations between a test and the job elements. The use of test experts is illustrated by Schmidt and his colleagues (Hirsh, Schmidt, & Hunter, 1986; Schmidt, Hunter, Croll, & McKenzie, 1983) who had personnel psychologists estimate the numerical value of test validities. In those instances when test experts are not available to an organization, internal options such as relating incumbent or supervisory ratings of performance on the job elements to test performance, would be more practical. In support of incumbent ratings, Primoff (1971) showed that self-ratings of abilities by job applicants were correlated greatest with tests measuring those abilities, and his data indicated that J-coefficients computed with the B^* component derived from these ratings were highly related ($r = .97$) to test intercorrelations.

The present study compared several procedures for estimating the component relations that define the J-coefficient. Component relations were estimated using supervisory and incumbent performance ratings, policy capturing, importance ratings, and test expert ratings. Specifically, the accuracy of the procedures was evaluated by comparing the resulting J-coefficients to criterion-related validities that were obtained with a concurrent validation design.

Method

Organizational Setting and Participants

The study was conducted in a branch plant (484 tenured employees) of a medium-sized, international manufacturer of precision automotive

accessories. The participants were 53 machinists and their 4 department supervisors. Participants volunteered from the drill, mill, grinder, and lathe departments. Participants were told that they would be providing data for a research project and that each person's data would be held in anonymity. Two other occupations were also examined, but results are not reported because of small sample sizes and incomplete data.

The organization used a battery of 14 predictors that was followed by interviews to make selection decisions. The predictors were the Otis Quick-Scoring Mental Ability Tests (Otis, 1965), the seven scales of the Humm-Wadsworth Temperament Scale (Humm & Wadsworth, 1935), the Bennett Mechanical Comprehension Test (Bennett et al., 1980), the Revised Minnesota Paper Form Board Test (Likert & Quasha, 1970), and the three scales and total score of the MacQuarrie Test for Mechanical Ability (MacQuarrie, 1953). Predictor scores were available from personnel records for validating these tests. Participants were unaware of predictor scores and test validities.

The organization conducts yearly evaluations of its employees. The employees are ranked and rated by their supervisors and co-workers according to their value to the department. The member evaluation committee uses both sources of information to produce an overall ranking of the 484 employees. The committee is composed of permanent members (department supervisors and upper-level management), variable members (supervisors are chosen to represent their work area on a 1-year basis), and temporary members (each supervisor may become a member to clarify or provide more information about a specific subordinate). First, the committee uses all available information (including supervisory and co-worker evaluations) to produce a reranking of employees according to their value to their department. Next, the committee ranks all employees according to their worth to the company. For this overall ranking, the committee considers the contribution of the department to the company, as well as other skills of the employee that are important to the company but not necessarily to the department (e.g., the employee may be able to work in other departments or translate foreign orders). The overall ranking of employees by the committee was completed 1 month prior to the study; no ties were permitted. The ranks of the 53 machinists were transformed to z scores for analysis; the equation for transformation was based on the 53 cases.

Job Analysis

A five-step procedure was used to identify the determinants of job performance (see Hamilton, 1977). First, 12 job incumbents were asked to independently write five effective and five ineffective critical incidents. Following the generation of the incidents, groups of 3 incumbents met in conferences to discuss and clarify their incidents, as well as to generate additional incidents (Campbell, Dunnette, Arvey, & Hellervik, 1973). Third, the authors edited all incidents for redundancies and ambiguities, analyzed the content of the remaining incidents, and generated six job element titles and definitions. The job elements and numbers of incidents were craftsmanship ($n = 17$), alertness ($n = 15$), practical intelligence ($n = 14$), planning and organization ($n = 11$), and positive work attitude ($n = 17$). Next, a conference was held with the four supervisors to clarify the job elements and their behavioral incidents and to verify that the job elements covered all aspects of job performance. The supervisors generated the additional job element of safety and 17 critical incidents for it. Finally, the supervisors and 2 incumbents scaled the incidents for each job element according to a method developed by Taylor (1968). Four to six incidents were selected and placed on 10-point scales in an example-anchored format for each job element. These job element scales were subsequently used for obtaining ratings from job experts.

Procedure

J-coefficients were calculated with components that were obtained from several sources of judgments. Table 1 contains a summary of the

Table 1
Combinations of Sources Used to Generate the J-Coefficients

Combination	Numerator ^a						
	Equation 1		Equation 2		Denominator		
	B_X^*	R_{YE}	B_Y^*	R_{XE}	R_{XE}	R_{EE}^{-1}	R_{YE}
1	S	P	—	—	S	S	P
2	—	—	P	S	S	S	P
3	S	P	—	—	S	P	P
4	—	—	P	S	S	P	P
5	I	P	—	—	I	I	P
6	—	—	P	I	I	I	P
7	I	P	—	—	I	P	P
8	—	—	P	I	I	P	P
9	T, P ^b	P	P	T	T	P	P
10	S	Im	Im, S ^b	S	S	S	Im
11	I	Im	Im, I ^b	I	I	I	Im
12	S	C	—	—	S	C	C
13	—	—	C	S	S	C	C
14	I	C	—	—	I	C	C
15	—	—	C	I	I	C	C
16	T, S ^b	S	S	E	T	S	S
17	T, I ^b	I	I	T	T	I	I

Note. S = supervisor; I = incumbent (self); P = policy capturing; Im = importance; C = co-worker; and T = test expert. B_X^* = column matrix of standard score regression coefficients from the regression of test performance on the job elements; B_Y^* = column matrix of standard score regression coefficients from the regression of job performance on the job elements; R_{XE} = column matrix of correlations between test performance and the elements; R_{YE} = column matrix of correlations between job performance and the job elements; and R_{EE}^{-1} = inverse matrix of correlations among job elements.

^a Source entries in the table for the numerator of both equations indicate that the J-coefficient could be calculated using either equation in that combination.

^b The notation specifies the letter before the comma as the source for the correlations of the job elements with test or job performance, R_{XE} , or R_{YE} , and after the comma as the source for the inverse matrix of correlations among the job elements, R_{EE}^{-1} .

combinations of the sources that were used to generate 17 different J-coefficients. Of course, Equations 1 and 2 produced numerically equal estimates when the same sources were employed.

Supervisory and incumbent (self-) ratings. Job incumbents rated themselves, and their supervisors rated them, on each of the job elements and on overall job performance. Ratings on the job elements were made using the example-anchored format described earlier. Ratings of overall job performance were made on a 10-point scale. This scale was anchored verbally with *extremely low*, *average*, and *extremely high*. Data from these two sources provided estimates of B_X^* and R_{YE} for Equation 1, B_Y^* and R_{XE} for Equation 2, and the inverse correlation matrix among job elements, R_{EE}^{-1} .

Co-worker ratings. A total of 32 job incumbents were asked to complete a modified policy capturing task. They rated co-workers on each of the job elements, and then on overall job performance. Raters were instructed to think of an actual or hypothetical co-worker in each of three categories (*most*, *typical or average*, and *least preferred* co-worker) and to rate them. Raters were cautioned *not* to rate co-workers by giving them maximum or minimum ratings but to rate "realistically." The co-worker ratings were combined to provide an estimate of R_{YE} for Equation 1, B_Y^* for Equation 2, and R_{EE}^{-1} .

Policy capturing. Job element intercorrelation matrices from the supervisor, self-, and co-worker ratings were averaged to define a theoreti-

cal correlation matrix (Naylor & Wherry, 1965). The average matrix was used in a profile generation program (Dickinson & Wherry, 1973) to produce 100 profiles of hypothetical workers. They were checked by two supervisors for their representativeness of actual incumbents, and all profiles were judged to be acceptable. Booklets of the profiles were distributed to 10 job incumbents along with the example-anchored rating scales. These incumbents were instructed to rate the overall job performance of the hypothetical employees.

Each incumbent's ratings were regressed on the job element scores to obtain a rating policy, B_Y^* . The rating policies of the incumbents were clustered to derive a composite policy (Christal, 1963). The policies of the 10 incumbents were quite homogeneous and adequately described by a single, composite policy. The squared multiple correlation before clustering the policies was .93, and dropped only to .91 after clustering. The composite policy provided an estimate of B_Y^* . In addition, the intercorrelations among the profiles of job element scores were used to derive an estimate of R_{EE} , and R_{YE} was estimated from the matrix product $R_{EE} B_Y^*$.

Importance ratings. All supervisors and the incumbents who did not make co-worker ratings were asked to rate the importance of the job elements to job performance. Importance ratings were made on a 10-point scale, *very low* (1) and *very high importance* (10); these were the only verbal anchors. The incumbent and supervisor ratings were averaged for each element, and the obtained averages were divided by 10. These values defined an estimate of R_{YE} . This estimate was used in combination with R_{EE} issuing from supervisor and incumbent sources to estimate B_Y^* from the matrix product $R_{EE}^{-1} R_{YE}$.

Test expert ratings. One faculty member and four doctoral-level students from an industrial/organizational psychology program estimated the correlations between tests and job elements. Judges read definitions of the tests and job elements, and then rated the degree of correlation between each test and job element on a scale ranging from -1.00 to 1.00. The ratings were averaged to estimate R_{XE} and used in combination with R_{EE} matrices from the policy capturing, supervisor, and incumbent sources to estimate B_X^* from the matrix product $R_{EE}^{-1} R_{XE}$.

Results

J-coefficients were computed with the 17 combinations of sources described in Table 1 for each of the 14 predictors. Three validities were computed for each predictor with a concurrent validation design. These validities used as criteria the overall ranking of worth to the company by the member evaluation committee and the ratings of overall job performance by supervisors and incumbents. The validities ranged in value from -.21 to .29. For each combination of sources and a criterion, the 14 J-coefficients and validities were correlated and a mean difference was computed. The correlations and mean differences were used to assess the accuracy of the J-coefficients that were generated by the 17 combinations of sources.

Correlations Between J-Coefficients and Validities

The correlations between the J-coefficients and validity coefficients were evaluated using the *t* test of zero correlation. Although the assumptions of this test were violated to an unknown degree in this research, Monte Carlo simulations (Edgell & Noon, 1984) suggest that the test is robust and can be used to assess correlations of the magnitude reported in Table 2 at traditional levels of significance. The obtained correlations indicate that several combinations of sources were able to generate J-coefficients that reproduced the pattern of validities. The

Table 2
Correlations and Mean Differences Between J-Coefficients and the Member Evaluation Committee, Supervisor, and Incumbent Validities

Combination ^a	Correlation			Mean Differences ^b		
	Committee	Supervisor	Incumbent	Committee	Supervisor	Incumbent
1	.609**	.940**	-.333	.069	.011	.125
2	.644**	.984**	-.327	.061	.003	.117
3	.631**	.913**	-.383	.077	.019	.134
4	.665**	.964**	-.394	.066	.009	.123
5	-.277	-.087	.735**	-.089	-.146	-.032
6	-.250	-.158	.927**	-.072	-.130	-.016
7	-.314	-.112	.757**	-.069	-.127	-.013
8	-.296	-.203	.939**	-.053	-.111	-.004
9	-.064	.027	-.027	-.061	-.119	-.004
10	.548*	.868**	-.290	.071	.014	.128
11	-.198	-.072	.773**	-.098	-.156	-.041
12	.457*	.769**	-.276	.007	-.051	.064
13	.678**	.965**	-.309	.062	.004	.119
14	-.314	-.119	.683**	-.042	-.100	.014
15	-.424	-.219	.845**	.003	-.055	.059
16	-.051	.030	.094	-.084	-.142	-.028
17	-.149	-.100	.021	-.094	-.151	-.037

^a The sources for the combinations are defined in Table 1.
^b A positive mean difference indicates that the validities were on the average greater than the J-coefficients.
* $p < .05$. ** $p < .01$.

results depend on the type of validities and the source for estimating B^*_X and R_{XE} . The committee and supervisory validities correlated significantly only with J-coefficients whose source for B^*_X and R_{XE} included supervisory ratings of incumbents on the job elements. In contrast, incumbent validities correlated significantly only with J-coefficients whose sources included self-ratings on the job elements. For all types of validities, policy capturing, importance, and co-worker sources were useful for estimating B^*_Y or R_{YE} .

Mean Differences Between J-Coefficients and Validities

J-coefficients and validities were compared to assess their mean differences. Both independent and correlated t tests were computed, because the data did not clearly meet the assumptions of either test. As shown in Table 2, none of the mean differences were statistically significant by either test. Because violation of the assumptions of these tests inflates Type I error, these results suggest that the J-coefficients tended to be unbiased estimators of the validities. Furthermore, these differences compare favorably to the amount of over- and underestimation ($\pm .09$) that Primoff (1959) reported for his research.

Discussion

This study demonstrated that several procedures can be used for estimating the component relations that define the J-coefficient. Several combinations of these procedures produced J-coefficients that were similar in pattern and mean value to criterion-related validities. These combinations suggest sources of judgments that can be used to generate J-coefficients. Supervisory and incumbent (self-) ratings were clearly superior to test expert ratings for estimating R_{XE} . None of the com-

binations that utilized our test expert ratings provided accurate estimates of the validities. These results disagree with those of Schmidt et al. (1983), who found that expert estimates of criterion-related validities were quite similar in absolute value compared to the actual validities. However, the judges used by Schmidt et al. were hand-picked experts in the field of personnel selection. In a subsequent study, Hirsh et al. (1986) found that recent PhDs made judgments that were much less accurate estimates of observed validities. Most of the experts in our study were graduate students who had less experience than the judges used by Schmidt and his colleagues. Apparently, test experts must have extensive experience in personnel selection before they can make accurate judgments of correlations. In comparison to test experts, supervisors and incumbents did not make direct judgments of correlations. Rather, R_{XE} was estimated empirically by the correlations of supervisory and self-ratings of job element performance with test scores available from personnel records. Furthermore, B^*_X was estimated empirically through the regression of these ratings on the test scores. This latter approach is analogous to Primoff's (1959) initial use of the J-coefficient. He obtained and updated B^*_X estimates through successive validation studies of his battery of tests for the 62 job elements. In the present study, B^*_X was also estimated solely from judgmental data. Test expert judgments of test-element correlations were combined with R_{EE}^{-1} matrices from policy capturing, supervisor, and incumbent sources. As previously discussed, our test expert ratings were a poor source for estimating components of the J-coefficient, such that none of procedures that were based solely on judgmental data yielded J-coefficients that were accurate estimators of validities. Nonetheless, a purely judgmental approach holds the unfulfilled promise of an "art of decision" in selection research.

Several procedures for obtaining judgments to estimate B_X^* and R_{XE} come to mind for future research. In a policy capturing task, experts could rate the test performance of hypothetical job incumbents from job element scores. The composite policy of judges would estimate B_X^* , and it could be used to produce an estimate of R_{XE} from the matrix product $R_{EE}B_X^*$. In contrast, the performance of hypothetical incumbents on each job element could be rated from their test scores. The regression of test scores on the ratings would be used to estimate B_X^* , and the correlations between the test scores and the ratings would be used to estimate R_{XE} . The latter policy-capturing procedure was used by Mullins and Usdin (1970) for scores on a battery of 10 tests to predict performance in four training courses. They found that correlations obtained with policy capturing underestimated empirical correlations only by an average of 2%. Finally, experts could rate the relevance of each test item for performance on each job element. This procedure is similar to rating the importance of each job element for job performance to estimate R_{YE} (Primoff, 1955, 1959). An averaging of relevance ratings would estimate the correlation between an item and a job element. Averages of all correlations between items and job elements along with the average intercorrelation among items, usually obtainable from test manuals, could be used in formulas for composite correlations (Ghiselli, Campbell, & Zedeck, 1981, p. 163) to estimate R_{XE} .

No source for estimating R_{YE} and B_X^* was found to be superior. Policy capturing, co-worker and importance ratings in combinations with other sources generated J-coefficients that produced accurate estimates. However, the correlations between the estimates and validities were significant only when performance on the job elements and the job were evaluated by the same source. This result can be interpreted either as a method bias in the collection of data or as differential perceptions of performance by supervisors and job incumbents. Until future research clarifies the result, criterion-related validities should be estimated with J-coefficients that are generated from the same source of expert judgments that would be used to compute the validities.

In the absence of research evidence to support the use of one source over the other for estimating the relations between job elements and job performance, the choice is left to more practical considerations: amount of preparation, the number of raters, and amount of time to make judgments. Policy capturing demanded the most preparation by the investigators (e.g., generation of the profile scores, checking their representativeness, and reproduction for their distribution to raters), and it also required the most time (100 ratings) of the raters. Nonetheless, policy capturing required the fewest (10) raters to obtain estimates of B_X^* and R_{YE} . Both co-worker and importance ratings required minimal preparation and time for administration. However, only the co-worker source produced estimates of B_X^* for Equation (2). Importance ratings must be used in combination with other sources that estimate both R_{XE} and B_X^* .

The findings of this research are limited by the nature of the tests and their validities. The tests cannot be considered representative of the predictors used in industrial settings (cf. Ghiselli, 1966; Guion, 1965), and the range of our test validities (-.21 to .29) did not cover the range to be expected (-.13 to .44) for a machinist occupation (cf. Ghiselli, 1966, 1973). However,

there was no string of zero validities to produce spurious correlations between the J-coefficients and the test validities. In addition, the consistency of our findings contraindicates that they were due to the nature of the validities. Policy capturing was consistently an effective method; the use of test experts, most of whom were graduate students, was consistently an ineffective method; and ratings were effective when used in conjunction with same source to estimate component relations.

Despite the potential of the J-coefficient, it has attracted little attention in the literature. The available research, however, suggests that it can be useful for personnel selection. Perhaps the lack of attention was because the seminal work on the J-coefficient used one job analytic method for identifying job elements and a particular battery of tests for estimating criterion-related validities (Primoff, 1964). The present study illustrates that these are not inherent limitations of the J-coefficient. Other job analytic methods can be used to produce J-coefficients that are accurate estimators of criterion-related validities for a variety of test predictors. This frees the selection researcher to generate the J-coefficient from job elements and tests that are more relevant and acceptable to the organization to better account for job success.

References

- Balma, M. J. (1959). The concept of synthetic validity. *Personnel Psychology*, 12, 395-396.
- Bennett, G. K., et al. (1980). *Mechanical comprehension test*. San Antonio, TX: Psychological Corporation.
- Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. (1973). The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology*, 57, 15-22.
- Christal, R. A. (1963). *JAN: A technique for analyzing individual and group judgment* (PRL-TDR-63-3, ASTIA Document AD-403-813). Lackland AFB, Texas: 6570th Personnel Research Laboratory, Aerospace Medical Division.
- Dickinson, T. L. (1976). *The J-coefficient: Statistical definitions and some procedures for its calculation* (Tech. Rep. No. 3). Ft. Collins: Colorado State University, Department of Psychology, Industrial Psychological Association of Colorado.
- Dickinson, T. L., & Wherry, R. J., Sr. (1973). A FORTRAN program for generating multiple samples of multivariate data with arbitrary population parameters. *Educational and Psychological Measurement*, 33, 715-718.
- Dickinson, T. L., & Wijting, J. P. (1976, May). *Policy capturing as a procedure for synthetic validation*. Paper presented at the meeting of the Rocky Mountain Psychological Association, Phoenix, AZ.
- Edgell, S. E., & Noon, S. M. (1984). Effect of violation of normality on the *t* test of the correlation coefficient. *Psychological Bulletin*, 95, 576-583.
- Ghiselli, E. E. (1966). *The validity of occupational aptitude tests*. New York: Wiley.
- Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. *Personnel Psychology*, 26, 461-477.
- Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco: Freeman.
- Guion, R. M. (1965). *Personnel testing*. New York: McGraw-Hill.
- Guion, R. M. (1976). Recruiting, selection, and job placement. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 777-828). Chicago: Rand McNally.
- Hamilton, J. W. (1977). *J-coefficient: A comparison of procedures*. Unpublished doctoral dissertation, Colorado State University.

- Hirsh, H. R., Schmidt, F. L., & Hunter, J. E. (1986). Estimation of employment validities by less experienced judges. *Personnel Psychology*, 39, 337-344.
- Humm, D. G., & Wadsworth, G. W. (1935). The Humm-Wadsworth Temperament Scale. *American Journal of Psychiatry*, 92, 163-200.
- Likert, R., & Quasha, W. H. (1970). *Revised Minnesota Paper Form Board Test*. San Antonio, TX: Psychological Corporation.
- MacQuarrie, T. W. (1953). *MacQuarrie Test for Mechanical Ability*. New York: CTB/McGraw-Hill.
- Mullins, C. J., & Usdin, E. (1970). *Estimation of validity in the absence of a criterion* (AFHRL-TR-70-36). Lackland Air Force Base, TX: Personnel Division.
- Naylor, J. C., & Wherry, R. J., Sr. (1965). The use of simulated stimuli and the "JAN" technique to capture and cluster the policies of raters. *Educational and Psychological Measurement*, 25, 969-986.
- Otis, A. S. (1965). *Otis quick-scoring mental ability tests*. New York: Harcourt, Brace, & World.
- Primoff, E. S. (1955). *Basic formulae for the J-coefficient* (Test Technical Series No. 25). Washington, DC: U.S. Civil Service Commission, Standards Division.
- Primoff, E. S. (1959). Empirical validations of the J-coefficient. *Personnel Psychology*, 12, 413-418.
- Primoff, E. S. (1964). *Test selection by job analysis: The J-coefficient* (2nd ed.; Test Technical Series No. 20). Washington, DC: U.S. Civil Service Commission, Personnel Measurement Research and Development Center.
- Primoff, E. S. (1971). *Preliminary report on the use of self-ratings to provide J-coefficient data*. Unpublished manuscript. U.S. Civil Service Commission, Standards Division.
- Schmidt, F. L., Hunter, J. E., Croll, P. R., & McKenzie, R. C. (1983). Estimation of employment test validities by expert judgment. *Journal of Applied Psychology*, 68, 590-601.
- Taylor, J. B. (1968). Rating scales as measures of clinical judgment: A method for increasing scale reliability and sensitivity. *Educational and Psychological Measurement*, 28, 747-766.
- Urry, V. W. (1978). *Some variations on derivations by Primoff and their extensions* (TN-78-3). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center.
- Wherry, R. J. (1955). *A review of the J-coefficient* (Assembled Test Technical Series No. 26). Washington, DC: U.S. Civil Service Commission Standards Division, Test Development Section.

Received February 6, 1986

Revision received July 3, 1986 ■

Low Publication Prices for APA Members and Affiliates

Keeping You Up-to-Date

All APA members (Fellows, Members, and Associates) receive — as part of their annual dues — subscriptions to the *American Psychologist*, the *APA Monitor*, and *Psychology Today*.

High School Teacher and Student Affiliates receive subscriptions to the *APA Monitor* and *Psychology Today*, and they can subscribe to the *American Psychologist* at a significantly reduced rate.

In addition, all members and affiliates are eligible for savings of up to 50% on other APA journals, as well as significant discounts on subscriptions from cooperating societies and publishers (e.g., the British Psychological Society, the American Sociological Association, and Human Sciences Press).

Essential Resources

APA members and affiliates receive **special rates** for purchases of APA books, including the *Publication Manual of the APA*, the *Master Lectures*, and *APA's Guide to Research Support*.

Other Benefits of Membership

Membership in APA also provides eligibility for low-cost insurance plans covering life; medical and income protection; hospital indemnity; accident and travel; Keogh retirement; office overhead; and student/school, professional, and liability.

For more information, write to American Psychological Association,
Membership Records, 1200 Seventeenth Street NW, Washington, DC 20036

Reactions to Procedural Injustice in Payment Distributions: Do the Means Justify the Ends?

Jerald Greenberg

Faculty of Management and Human Resources, Ohio State University

In a laboratory study, 192 undergraduate students performed a task for which they received either high, medium, or low monetary outcomes as a result of a fair or unfair procedure. Subjects reported that medium and high outcomes were fair regardless of the procedure used, but that low outcomes were only fair when they were based on a fair procedure. The outcomes received, however, had no impact on ratings of the fairness of the procedures used. These results corroborated earlier findings from the dispute-resolution literature, but extend them to reward-distribution contexts in which different manipulations of procedural justice were used. The limitations of equity theory in accounting for improprieties in organizational procedures is discussed.

It is well established that distributions of organizational rewards (such as pay raises, promotions, job status, and the like) influence, in some manner, the attitudes and behavior of employees (Lawler, 1977). Indeed, several theoretical conceptualizations of justice in organizations, most notably equity theory (Adams, 1965; Walster, Walster, & Berscheid, 1978), have focused extensively on how distributions of organizational rewards (also referred to as *outcomes*) affect job satisfaction and performance (for a review, see Greenberg, 1982). This legacy of theory and research, although it reveals a great deal about reactions to the nature and level of organizational rewards, provides little insight into possible effects caused by the manner in which these rewards are established. As a result, questions remain about whether or not (and if so, how) the *way* organizational rewards are determined influences reactions to them.

A reorientation in emphasis from *what* the rewards are to *how* they are determined follows from theoretical conceptualizations of *procedural justice* (e.g., Thibaut & Walker, 1975), offering a broader, more procedurally oriented conceptualization of justice than the traditionally outcome-oriented, *distributive justice* perspective of equity theory (see Thibaut & Walker, 1978). Whereas distributive justice focuses on the fairness of a distribution of resources, procedural justice focuses on the fairness of the procedures used to make those distributive decisions. Recent research on procedural justice has highlighted the importance of resource distribution procedures as determinants of fairness in organizations (for reviews, see Folger & Greenberg, 1985; Greenberg, 1986a; Greenberg & Folger, 1983). Alexander and Ruderman (in press), for example, surveying more than 2,800 federal employees, found that employees' concerns about how their salaries were determined accounted for more variance in job satisfaction than the level of

those salaries. Similarly, with respect to another organizational outcome, performance evaluations, Landy, Barnes-Farrell, and Cleveland (1980) found that the process by which workers' performance appraisals were determined was related to the perceived fairness of their evaluations, regardless of how positive or negative they were. More generally, in a survey of executives, Sheppard and Lewicki (in press) found that procedural factors (the way outcomes were determined) were reported as being more critical than specific outcome variables themselves as determinants of fair and unfair treatment in their organizations. Taken together, such evidence highlights the importance of procedural aspects of justice in the context of organizational reward distributions.

What makes a reward distribution procedure unfair? In his theory of procedural fairness, Leventhal (1976, 1980; Leventhal, Karuza, & Fry, 1980) posited that fair allocation procedures are characterized by resource distributions that are consistent across persons and over time, free from bias, based on accurate information, correctable, representative of all recipients' concerns, and based on prevailing moral and ethical standards. Recent survey studies have shown that workers report extreme dissatisfaction with resource distribution procedures that violate these standards, and believe them to be unfair. For example, in their survey study, Sheppard and Lewicki (in press) found that accounts of unfair incidents frequently alluded to elements in Leventhal's conceptualization (such as inconsistency and bias in reward allocations). Similarly, Greenberg (in press) found that workers reported that performance evaluations made without keeping accurate performance records were unfair. Finally, Barrett-Howard and Tyler (1986) have shown that role-playing subjects responded negatively to violations of Leventhal's procedural justice standards in imagined work settings. Together, this body of work suggests that criteria for stipulating procedurally fair and unfair practices exist that recently have begun to receive empirical validation.

With increasing awareness of the importance of procedural justice in organizations and preliminary demonstrations of adverse reactions to procedural justice violations comes the need to know how reactions to outcome distributions and the proce-

Support for this investigation was provided in part by Grant INT-8304375 from the National Science Foundation.

Correspondence concerning this article should be addressed to Jerald Greenberg, Faculty of Management and Human Resources, Ohio State University, 1775 College Road, Columbus, Ohio 43210-1399.

dures from which they are derived are related. The primary question of interest is how the fairness of the procedures used influences the perceived fairness of the resulting outcomes. Do fair procedures lead to fair outcomes? There is good reason to hypothesize that there would be a main effect of procedures on judgments of distributive fairness such that fair procedures lead to judgments of fairer outcomes than unfair procedures. Conceptual support for this hypothesis may be derived from several sources. For one, Leventhal (1976) claimed that unfair procedures cannot yield fair outcome distributions. Similarly, Thibaut and Walker (1975) suggested that there may be a perceptual overlap between distributive justice and procedural justice such that the fairer the procedure used to determine outcomes, the more those outcomes are likely to be evaluated as being distributively just. Indirect empirical support for the hypothesis can be found in studies simulating legal dispute resolution techniques in which the perceived fairness of the resulting verdicts were influenced by the procedures on which they were based; procedures promoting personal participation in adjudication were seen as fairest (Lind, Kurtz, Musante, Walker, & Thibaut, 1980; Walker, Lind, & Thibaut, 1979). Although these studies involve contexts and operational definition of procedural injustice that are quite different from the present study (and therefore only modest empirical justification for the hypothesis), they provide good insight into the question of interest.

A second major issue addressed in the present study is the opposite question—namely, how do the outcomes received influence the perceived fairness of the procedures by which they were determined? It already has been established that higher outcomes are seen as being fairer than lower outcomes—the so-called *egocentric bias* in perceptions of distributive justice (e.g., Greenberg, 1983a). We may ask whether this effect generalizes to perceptions of procedural justice as well? The results of previous studies reveal a mixed answer to this question. Although several investigations have reported that beneficiaries of positive outcomes tend to view the procedures leading to them as being fair (e.g., LaTour, 1978; Tyler & Caine, 1981, Studies 2 and 4), others have reported that outcome favorability does not influence perceptions of procedural justice (e.g., Lind et al., 1980). The evidence bearing on this question, is not only contradictory but is based on retrospective questionnaires or uninvolved role-playing techniques, which weakens our confidence in its generalizability. Accordingly, the present study manipulated outcome level and measured its immediate impact on perceptions of procedural fairness. Corroboration of the *egocentric bias* would require finding a significant main effect of outcome level in ratings of outcome fairness such that higher outcomes are seen as fairer than lower outcomes. To the extent that this effect generalizes to perceptions of procedural fairness, a similar main effect of outcome level on procedural fairness ratings would be expected.

Finally, the present study provided an opportunity to explore the possibility that reactions to procedural injustices would be moderated by their underlying causal basis—either individual or organizational. This variable is suggested by research showing that reactions to distributive justice differ as a function of whether the inequity is caused by an individual or an organization (Greenberg, 1986c; Leventhal, Younts, & Lund, 1972). The following question was asked: Would individually based or organiza-

nizationally based unfair procedures elicit greater reactions? A specific prediction is not immediately forthcoming due to recent evidence suggesting potentially opposite reactions. On one hand, Folger and Martin (in press) have shown that subjects are likely to react more strongly against agents of injustice who are expected to be able to continue their unjust actions in the future. On the other hand, evidence also suggests that victims of injustice may refrain from striking back at causal agents when they believe their actions will have little impact (Martin, Brickman, & Murray, 1984). Combining these two findings makes it difficult to predict how individuals will respond when they believe the procedures they confront are based on organizational policies, which may be expected to be more enduring and serious sources of injustice, but may also be more difficult to correct. The present study explored the possibility that attitudinal and behavioral reactions to procedural justice and injustice would be differentially influenced by the individual or organizational basis of their origin, although no specific hypotheses were tested.

Method

Subjects and Design

The participants were 192 undergraduate students (96 men and 96 women) at a midwestern university who volunteered to participate in a study allegedly concerned about “consumer use of sales catalogues.” In exchange for taking part in the study, subjects were promised a payment of “up to \$8” for the 1-hr session. (This phrasing of the stated payment amount was revealed in pilot testing to result in a potential range of payments perceived to be fair compensation for participating in the study.) Five additional subjects (3 men and 2 women), evenly distributed over the experimental conditions, also participated in the study, but their responses were not analyzed due to their failure to follow experimental instructions.

The overall design of the experiment was a $3 \times 2 \times 2 \times 2$ factorial in which the independent variables were outcome level (high, medium, or low), procedural fairness (fair or unfair), origin of procedure (individual or organizational), and sex of subject. There were an equal number of men and women randomly assigned to each cell.

Procedure

Pairs of same-sex subjects participated in each experimental session. They were told that they would be performing a “catalogue searching task,” and then would complete a brief questionnaire assessing their reactions to the task.

Experimental task. The purpose of the experimental task was to provide an apparent basis for the experimentally manipulated payments that followed. So as to avoid arousing subjects’ suspicions about the experimental manipulations, a task was used that has been shown in previous research to be one for which subjects have no preconceived standards of productivity (Greenberg, 1983b). The task consisted of locating specified items in a department store catalogue and copying their prices onto index cards on which the items were identified. The rationale was given that this study was being conducted to find out how the design of catalogues influences people’s ability to use them. After the experimenter demonstrated the task and answered subjects’ questions about how to perform it, the subjects were escorted to opposite ends of the same room and were seated at desks containing the index cards, pencils, and catalogue needed to perform the task. Because subjects were seated facing opposite directions, they could not see each other’s work.

After performing this task for 45 min, the experimenter entered the work room and instructed subjects to stop working. He then handed each subject a large manila envelope into which he instructed them to place all of their index cards, both the completed and incomplete ones. (This practice minimized subjects' opportunities to assess their relative inputs.) The experimenter announced that he would get the subjects started on the questionnaire and arrange for their payment, but that first they would have to leave the workroom so that another group of subjects could be brought in. Subjects were then asked to go to any one of three nearby rooms, labeled "Room A," "Room B," and "Room C," to complete the study. The experimenter explained that because these rooms were small and contained only one desk, only one subject should enter a room, and wait there for the experimenter to return. This procedure made it possible for the experimenter to independently manipulate the experimental conditions in each session.

Independent variable manipulations. Between 3 and 6 min later the experimenter returned to each subject's experimental room and announced how their pay was determined. This information constituted the procedural fairness manipulation. In the *fair-procedure* condition subjects were told that their pay was based on how well they performed on the catalogue-searching task relative to the other person. Better workers were said to receive a higher proportion of the \$8 than poorer workers. In the *unfair-procedure* condition subjects were told that their pay was determined by the room that they selected. Each of the three rooms, it was explained, had a predetermined amount of money associated with it that constituted the payment of the person selecting it.¹ This manipulation of procedural injustice—a seemingly arbitrary procedure for determining pay—is justified by its inclusion of several procedural elements specified by Leventhal (1980), most notably that the procedure violates usual payment allocation practices and that it fails to base allocations on accurate performance information.

Following this, the experimenter commented on the origin of the allocation procedure he just explained. In the *individual* condition, the experimenter said that it was he who personally decided how the payment division decision was to be made. In the *organizational* condition, the experimenter explained that the decision to divide the pay this way was made by the large company that sponsored the research.

Based on the announced procedure, the experimenter explained, he would now pay subjects their share of the \$8. In particular, outcome level was manipulated by telling subjects that they would be receiving either \$7 (in the high-outcome condition), \$4 (in the medium-outcome condition), or \$1 (in the low-outcome condition). To enhance the salience of this manipulation, the experimenter took out eight \$1 bills and handed the appropriate number to the subject.

Dependent measures. After subjects took the money, the experimenter handed them a booklet containing six questionnaire items. Using 9-point bipolar scales, subjects were asked to indicate the following: their perception of the fairness of the payment they received, the fairness of the procedure used to determine their payment (for both items, 1 = *extremely unfair*, 9 = *extremely fair*), their concern over the pay they received, their concern over how their pay was determined (for both items, 1 = *extremely unconcerned*, 9 = *extremely concerned*), their liking for the experimental task, and their liking for the experimenter (for both items, 1 = *extremely dislike*, 9 = *extremely like*). The four questions assessing concern and liking were considered supplementary measures designed to provide insight into the reasons underlying responses to the two fairness questions.

Subjects were assured of the anonymity of their responses. In support of this claim subjects were given a letter-size envelope into which they were to insert their completed questionnaires. These envelopes, it was explained, were to be inserted into a large folder tacked onto a bulletin board on the wall before leaving the room. The folder was labeled "Catalogue Study—Completed Questionnaires" and contained some al-

ready-stuffed envelopes, thereby supporting the illusion of response anonymity.

On the bulletin board immediately to the left of the folder for depositing the completed questionnaires was a prominent notice on which there was printed an octagonal-shaped sign containing the words "STOP UNFAIR EXPERIMENTS." The remainder of the text read as follows, "Treated Unfairly in an Experiment? Call the Ethical Responsibility Board to Report any Unfair Treatment in Human Experimentation." At the bottom of the sign appeared the words "Take Our Number," immediately over a series of vertical cuts in the paper approximately 38-mm long. On each strip appeared a local telephone number that subjects could easily take by tearing one strip off the sign. To invite subjects to do so, four strips were already torn off the sign, leaving five phone-number strips for subjects. If one or more additional strips were missing after subjects left the room, that was taken as the subject's behavioral intention to express dissatisfaction with the experiment, providing a behavioral measure of dissatisfaction to supplement the questionnaire responses.

Debriefing. As subjects left their rooms they were intercepted by the experimenter, who conducted the postexperimental debriefing. Subjects were very carefully debriefed and were given the difference between what they were already paid and the maximum stated payment, thereby leaving each subject with a total of \$8. In postexperimental interviews no subjects expressed any suspicions about the actual purpose of the experiment or admitted having prior knowledge about it.

Results

Preliminary Analyses

A $3 \times 2 \times 2 \times 2$ multivariate analysis of variance was performed, using as between-subjects factors the three manipulated variables (outcome level, procedural fairness, and origin of procedure) and as an exploratory variable, sex of subject. The six questionnaire measures constituted the dependent variables. Statistically significant effects were obtained for outcome level, procedural fairness, and the interaction between them (all values of multivariate $F \geq 11.19$, $p < .001$). All other sources of variance were nonsignificant (all values of multivariate $F < 1.00$). Correlations between responses to the six questionnaire items were all nonsignificant; using Fisher's transformation, the mean of the 15 zero-order correlations in the matrix was .04, *ns*. Accordingly, none of the measures were combined prior to data analyses. On the basis of these findings, all analyses of the questionnaire results reported in this article are based on 3×2 (Outcome Level \times Procedural Fairness) univariate analyses of variance.

Perceived Fairness Measures

Analyses of outcome fairness revealed a significant main effect of outcome level, $F(2, 186) = 6.06$, $p < .005$, $\omega^2 = .08$.

¹ A "pre-inquiry quasi control group" (Orne, 1969) of 28 subjects from the same population as that used in the main study was used to assess the validity of this manipulation, without possible contamination created by knowledge of the outcomes. These participants were subjected to the same experimental procedure as the regular subjects, but they did not receive information about their outcome level or the origin of the procedure. The subjects (one half of the group) who received the fair-procedure manipulation ($M = 7.81$) reported that it was significantly fairer than did the other half who received the unfair-procedure manipulation ($M = 1.89$), $F(1, 26) = 89.76$, $p < .001$.

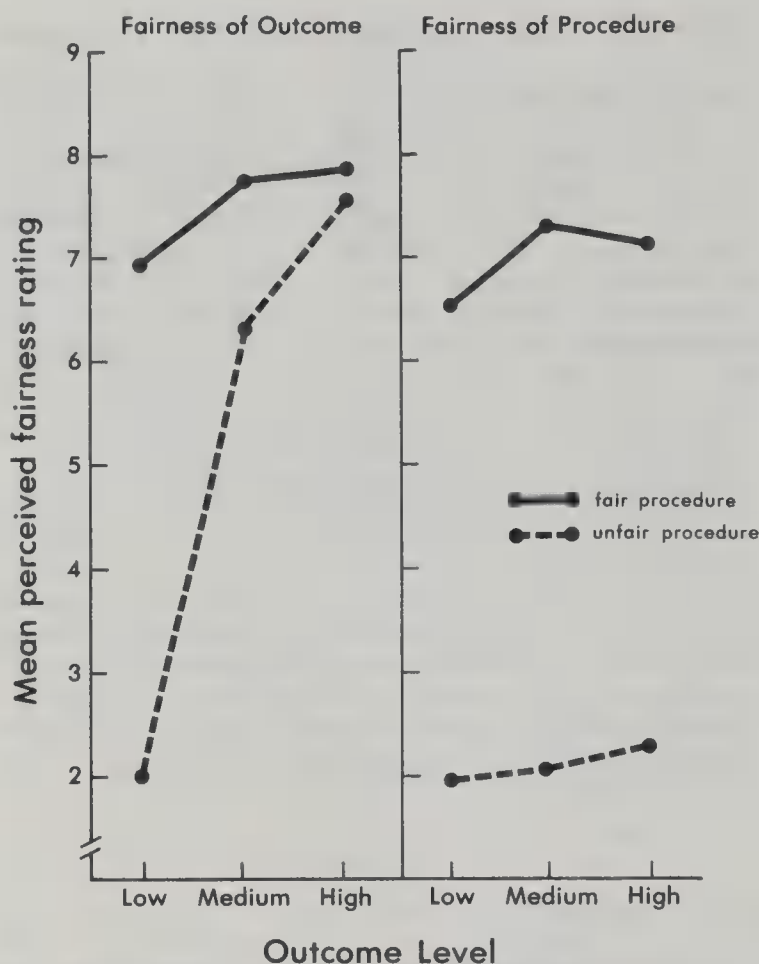


Figure 1. Mean ratings of outcome fairness and procedural fairness as a function of outcome level and procedure.

Newman-Keuls tests ($\alpha = .05$) revealed that this effect resulted from the tendency for subjects to report that low outcomes ($M = 4.45$) were less fair than medium outcomes ($M = 7.08$) or high outcomes ($M = 7.70$), which were not significantly different from each other. Also significant was the main effect of procedural fairness, $F(1, 186) = 12.18, p < .001, \omega^2 = .17$, such that fair procedures ($M = 7.55$) were believed to result in fairer outcomes than unfair procedures ($M = 5.27$). Both these effects, however, were qualified by a significant interaction between outcome level and procedural fairness, $F(2, 186) = 19.47, p < .001, \omega^2 = .27$. The means corresponding to this interaction are displayed in the left-hand panel of Figure 1.

Simple main effects tests across procedures revealed that a significant difference in the perceived fairness of outcomes occurred only at the low-outcome level, $F(1, 62) = 9.75, p < .001$, where fair procedures were seen as responsible for fairer outcomes than unfair procedures. This is in contrast to medium- and high-outcome levels, in which the procedures used were found to have no significant influence on the fairness of the resulting outcome; in both cases, $F < 2.00, ns$.

For ratings of the fairness of the procedures used, the only significant source of variance was the main effect of procedural fairness, $F(1, 186) = 20.69, p < .001, \omega^2 = .12$. As shown on the right side of Figure 1, the fair procedure was seen as being fairer than the unfair procedure. Unlike ratings of outcome fairness, this effect was not qualified by outcome level; for the interaction, $F < 1.00, ns$.

Concern Over Outcomes and Procedures

Analyses of subjects' ratings of amount of concern over the outcomes they received yielded as the only significant source of variance a main effect of outcome level, $F(2, 186) = 8.74, p < .005, \omega^2 = .19$. Newman-Keuls tests ($\alpha = .05$) revealed that this effect resulted from the tendency for subjects in the low-outcome condition ($M = 7.40$) to express a significantly higher degree of concern over outcomes than either those in the medium-outcome condition ($M = 3.75$) or in the high-outcome condition ($M = 4.15$), which were not significantly different from each other.

Analyses of subjects' concern over the procedures used resulted in a significant main effect of procedural fairness, $F(1, 186) = 4.12, p < .05, \omega^2 = .06$, such that higher concern was expressed under unfair conditions ($M = 5.62$) than fair conditions ($M = 3.10$). The meaning of this effect was qualified, however, by a significant Outcome Level \times Procedural Fairness interaction, $F(2, 186) = 16.73, p < .001, \omega^2 = .26$. Simple main effects tests were used to compare responses across procedures at each level of outcome. It was found that when outcome levels were low, subjects expressed significantly more concern over unfair procedures ($M = 8.05$) than fair procedures ($M = 2.80$), $F(1, 62) = 88.30, p < .001$. However, when outcome levels were either moderate ($M = 3.75$) or high ($M = 3.90$), there were no significant differences found between procedures; in both cases, $F < 1, ns$.

Liking for Task and Experimenter

The only significant source of variance found for ratings of liking for the task was the main effect of outcome, $F(2, 186) = 9.40, p < .001, \omega^2 = .15$. Newman-Keuls tests ($\alpha = .05$) revealed that the means across all three levels of outcome (low $M = 3.75$, medium $M = 5.53$, high $M = 7.28$) were significantly different from each other. The higher the outcome, the more the task was liked.

For ratings of liking for the experimenter, the only significant source of variance was the main effect of procedural fairness, $F(1, 186) = 25.61, p < .001, \omega^2 = .11$. This effect was the result of the tendency for subjects to show greater liking for experimenters who treated them fairly ($M = 6.75$) than those who treated them unfairly ($M = 2.82$).

Behavioral Intention Measure

The final dependent measure was the number of subjects who took the telephone number from the notice posted on the bulletin board—a measure of intention to complain about unfair treatment. It was found that only subjects in the unfair-procedure/low-outcome condition took the telephone number. Specifically 14 out of 32 subjects in this condition (43.75%) took the number, whereas no subjects in any of the other conditions did so.

Of the 14 subjects who took the telephone number, 12 were in the condition in which they were told that the procedure was the result of an organizational policy, whereas only 2 were in the condition in which they were told that the procedure was the result of an individual decision. The difference between the

proportion of subjects in the two cells who took the telephone number (12 out of 16 [75%] vs. 2 out of 16 [12.5%]) was statistically significant ($z = 3.57, p = .0023$).

Discussion

The present research addresses two principal questions—one about the effects of procedures on outcomes, and the complementary question about the effects of outcomes on procedures.

The Influence of Procedures on Outcomes

How do the procedures used affect the reactions to the outcomes received? Extrapolating from studies simulating legal dispute-resolution contexts, (e.g., Lind et al., 1980; Walker et al., 1979) it was hypothesized that fair procedures would lead to outcomes believed to be fairer than would those resulting from unfair procedures. Although this hypothesis was strongly supported, the relative perceived fairness of the rewards based on fair procedures compared to unfair procedures only manifested itself when outcomes were low. In contrast, medium- and high-level outcomes were reported to be equally fair regardless of the fairness of the procedures used to bring them about. Stated differently, the means (procedures used) justified the ends (outcomes received) only when those ends were positive (medium- or high-outcome levels). This pattern of results partially supports but qualifies Leventhal's (1976) claim that "procedural fairness is a necessary precondition for the establishment and maintenance of distributive fairness" (p. 230). The present findings suggest that procedural justice may be a necessary precondition for distributive justice, but only when the outcomes are low.

The way in which outcome level qualified the effects of procedure suggests that procedures may matter most to people when they result in negative outcomes. This possibility is supported by the present finding that expressed concern over procedures was highest when unfair procedures resulted in low payments. Not surprisingly, it was precisely under these conditions that subjects behaviorally expressed their concern over unfair treatment. Subjects took action in response to the procedural impropriety only when their outcomes were low, despite the fact that unfair procedures were seen as being equally unfair across all outcome levels. Why didn't subjects experiencing medium- and high-level outcomes react to the procedural injustice? The fact that they perceived these higher outcomes as being fair appears to have removed subjects' justification for taking action in response to unfair procedures. This interpretation of the results is bolstered by evidence that segments of society benefiting from unfair procedures may express concern over the unfair situation but refrain from doing anything to jeopardize their privileged position (Cohen, 1986).

It is interesting that victims of unfair procedures were more likely to take action directed at redressing the injustice when they believed the unfair procedure followed from an organizational policy than when it was an individual decision. An explanation for this difference is suggested by Folger and Martin's (in press) recent findings suggesting that reactions to procedural infractions are exaggerated when the infraction is especially serious and expected to continue in the future. To the extent that

the organizationally based procedural injustices were seen as being more serious and permanent than the individually based ones, then the present results are not surprising.²

Not surprisingly, subjects' positive reaction to their pay generalized to the task itself; subjects liked tasks better for which they were highly paid than those for which they were poorly paid. The fact that higher outcomes were believed to be fairer than lower outcomes supports previous research (e.g., Greenberg, 1983a) showing that subjects believed to be fair those outcomes that benefited themselves. This result is also consistent with Thibaut and Walker's (1975) finding that it is the loser of trials who are most likely to view the verdicts as being unfair. Despite the corroborative nature of the present evidence, caution is needed in interpreting and generalizing from it because the meaning of apparent improprieties may be challenged in the context of a laboratory experiment. Similarly, limitations imposed by the single-item dependent measures used in this study (e.g., possible ambiguities) also restrict the potential generalizability of the findings.

The Influence of Outcomes on Procedures

A complementary question of interest was how do the outcomes received influence reactions to the procedure used? The results showed that fair procedures were believed to be fairer than unfair procedures regardless of the resulting level of outcome. The egocentric bias found in ratings of outcome fairness (i.e., that higher outcomes are judged fairer than lower outcomes) did not generalize to ratings of procedural fairness. Of particular interest is the finding that even procedures leading to low outcomes were believed to be fair when they resulted from fair procedures. Analogous evidence has been obtained by Tyler (1984; Tyler & Folger, 1980), who found that citizens who were found guilty of a misdemeanor by a judge or who were cited for a traffic violation by a police officer believed they were treated fairly when the authority figure adhered to certain expected practices. Not surprisingly, such authority figures tended to be liked, as was the experimenter in the present study, when they followed a fair procedure.

Although outcome level did not influence judgments of the fairness of the procedure used, it did have other effects. For example, low outcomes aroused concern over outcomes and over the procedures by which they were obtained (especially when they resulted from unfair procedures), and diminished liking

² Indeed, the questionnaire responses of 127 pilot subjects supports the claim that injustices caused by organizational policy are more serious and more permanent than those caused by an individual decision. These subjects read vignettes describing cases in which low outcomes resulted from an individually based or organizationally based procedure calling for outcomes to be based on an arbitrary decision rule—the choice of a room. It was found that ratings of seriousness of the infraction and permanence of the procedure (both on 9-point scales, with higher ratings reflecting greater degrees of seriousness and permanence) were significantly higher in the organizationally based condition (M for seriousness = 6.74, and M for permanence = 7.12) than in the individual-decision condition (M for seriousness = 3.82, and M for permanence = 4.02); values of $F(1, 126) = 12.73$ and 14.16, respectively, in both cases $p < .001$.

for the task. These findings clearly discount the possibility that outcome level failed to influence procedural fairness judgments because it was not made salient in the study. Instead, although the effects of outcome level were recognized, they did not influence perceptions of the fairness of the procedures used. That is, the organizational ends received (monetary rewards) did not justify the means (procedures) by which they were attained.

A similar tendency for outcome favorability to have no influence on perceptions of procedural justice also has been found in a study by Lind et al. (1980), conducted in a legal dispute-resolution context. Comparisons between the present findings and those of Lind et al. must be made cautiously because of the very different experimental settings and operational definitions of procedural injustice used. Lind et al. operationally defined procedural injustice by limiting litigants' input into the decision-making process (inspired by Thibaut & Walker, 1975), whereas in the present study unfair procedures were created by using a capriciously applied allocation rule (inspired by Leventhal, 1980). Both procedures were believed to be unfair, seemingly because both are counternormative in their respective settings. Yet, it remains an untested possibility that different sources of procedural improprieties would have resulted in different reactions.

Implications

The present study has important implications for research and theory on procedural justice. Primarily, it shows that many of the same reactions to the absence of control over the decision-making process demonstrated in studies of dispute resolution also manifest themselves in response to the capricious reward allocation procedure used in the present experiment. In particular, the present results corroborated Lind et al.'s (1980) findings about the positive effects of outcome level on perceptions of distributive justice and the lack of impact of outcome level on perceptions of procedural justice. In addition, by showing the effects of manipulations of procedural fairness along the lines suggested by Leventhal (1980), the present study supports and extends preliminary investigations (e.g., Sheppard & Lewicki, in press) showing that such procedural concerns are expressed among workers by showing how these factors operate. However, because a compound operationalization of an unfair procedure was used in the present study (violating prevailing standards and failing to use accurate performance information), it is unclear precisely which procedural characteristics accounted for the results. Implications for organizational behavior research also are suggested by the present work. Notable in this regard is the suggestion that theoretical conceptualizations focusing on organizational rewards, such as equity theory (Adams, 1965) and expectancy theory (e.g., Porter & Lawler, 1968), may need to be expanded to incorporate considerations of *how* outcomes are determined as well as *what* they are. By focusing on relative outcome and input levels, equity theory is not equipped to interpret the observed differences in the perceived fairness of low-level rewards derived from fair and unfair procedures. To the extent that procedures qualify the meaning of outcomes (and reactions to them), as demonstrated, then it would be essential for conceptualizations of justice in organizations to incorporate procedural variables. Indeed, given that or-

ganizational procedures are more frequently cited than outcomes as causes of unfairness in organizations (Greenberg, 1986a; Sheppard & Lewicki, in press), and because procedures contribute more to job satisfaction than do outcomes (Alexander & Ruderman, in press), the need to explore organizational procedures is further emphasized.

The results of the present study suggest several potentially fruitful new directions for future research. Among these is the important unaddressed issue of how workers' reactions to procedural injustice influences their job performance. Although equity theory (e.g., Adams, 1965) explains how workers are likely to react to unfair outcomes, it remains unclear how these reactions may be qualified by unfair organizational procedures. In a related vein, it would appear useful for organizational researchers to assess the generalizability of the present findings by examining the effects of a variety of potentially important procedural variables (such as those suggested by Folger & Greenberg, 1985; Sheppard & Lewicki, in press). The restricted procedural improprieties examined in the present study, although responsible for interesting findings, may be of limited usefulness in permitting generalizations to be drawn about the general theoretical properties of procedural justice (Leventhal, 1980). Finally, a further potential limitation of the study rests in the fact that the laboratory methodology used precludes the possibility of directly generalizing from the present findings to field applications. However, as is the case in theoretically based research, the issue of generalizability applies to the phenomenon under investigation rather than to the research findings themselves. As future researchers begin to recognize the importance of distinguishing between the outcomes of managerial decisions and the procedures that led to these decisions in organizational contexts—as is just beginning to be done in the field of performance appraisal (Greenberg, 1986b)—it will become possible to assess the external validity of the concepts uncovered here.

References

- Adams, J. S. (1965). Inequity in social exchange. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 267–299). New York: Academic Press.
- Alexander, S., & Ruderman, M. (in press). The role of procedural and distributive justice in organizational behavior. *Social Justice Review*.
- Barrett-Howard, E., & Tyler, T. R. (1986). Procedural justice as a criterion in allocation decisions. *Journal of Personality and Social Psychology*, 50, 296–304.
- Cohen, R. L. (1986). Power and justice in intergroup relations. In H. W. Bierhoff, R. L. Cohen, & J. Greenberg (Eds.), *Justice in social relations* (pp. 65–84). New York: Plenum Press.
- Folger, R., & Greenberg, J. (1985). Procedural justice: An interpretive analysis of personnel systems. In K. Rowland & G. Ferris (Eds.), *Research in personnel and human resources management* (Vol. 3; pp. 141–183). Greenwich, CT: JAI Press.
- Folger, R., & Martin, C. (in press). Relative deprivation and referent cognitions: Distributive and procedural justice effects. *Journal of Experimental Social Psychology*.
- Greenberg, J. (1982). Approaching equity and avoiding inequity in groups and organizations. In J. Greenberg & R. L. Cohen (Eds.), *Equity and justice in social behavior* (pp. 389–435). New York: Academic Press.
- Greenberg, J. (1983a). Overcoming egocentric bias in perceived fairness through self-awareness. *Social Psychology Quarterly*, 46, 152–156.

- Greenberg, J. (1983b). Self-image versus impression management in adherence to distributive justice standards: The influence of self-awareness and self-consciousness. *Journal of Personality and Social Psychology*, 44, 5-19.
- Greenberg, J. (1986a). Organizational performance appraisal procedures: What makes them fair? In M. H. Bazerman, R. J. Lewicki, & B. H. Sheppard (Eds.), *Research on negotiation in organizations* (Vol. 1, pp. 25-41). Greenwich, CT: JAI Press.
- Greenberg, J. (1986b). Determinants of perceived fairness of performance evaluations. *Journal of Applied Psychology*, 71, 340-342.
- Greenberg, J. (1986c). Differential intolerance for inequity from organizational and individual agents. *Journal of Applied Social Psychology*, 16, 191-196.
- Greenberg, J. (in press). Using diaries to promote procedural justice in performance appraisals. *Social Justice Review*.
- Greenberg, J. & Folger, R. (1983). Procedural justice, participation, and the fair process effect in groups and organizations. In P. B. Paulus (Ed.), *Basic group processes* (pp. 235-256). New York: Springer-Verlag.
- Handy, F. J., Barnes-Farrell, J., & Cleveland, J. N. (1980). Perceived fairness and accuracy of performance evaluation: A follow-up. *Journal of Applied Psychology*, 65, 355-356.
- Latour, S. (1978). Determinants of participant and observer satisfaction with adversary and inquisitorial modes of adjudication. *Journal of Personality and Social Psychology*, 36, 1531-1545.
- Lawler, E. E. III (1977). Reward systems. In J. R. Hackman & J. L. Suttle (Eds.), *Improving life at work* (pp. 163-226). Santa Monica, CA: Goodyear.
- Leventhal, G. S. (1976). Fairness in social relationships. In J. W. Thibaut, J. T. Spence, & R. C. Carson (Eds.), *Contemporary topics in social psychology* (pp. 211-239). Morristown, NJ: General Learning Press.
- Leventhal, G. S. (1980). What should be done with equity theory? In K. J. Gergen, M. S. Greenberg, & R. H. Willis (Eds.), *Social exchange: Advances in theory and research* (pp. 27-55). New York: Plenum Press.
- Leventhal, G. S., Karuza, J., & Fry, W. R. (1980). Beyond fairness: A theory of allocation preferences. In G. Mikula (Ed.), *Justice and social interaction* (pp. 167-218). New York: Springer-Verlag.
- Leventhal, G. S., Younts, C. M., & Lund, A. K. (1972). Tolerance for inequity in buyer-seller relationships. *Journal of Applied Social Psychology*, 2, 308-318.
- Lind, E. A., Kurtz, S., Musante, L., Walker, L., & Thibaut, J. W. (1980). Procedure and outcome effects on reactions to adjudicated resolution of conflicts of interest. *Journal of Personality and Social Psychology*, 39, 643-653.
- Martin, J., Brickman, P., & Murray, A. (1984). Moral outrage and pragmatism: Explanations for collective action. *Journal of Experimental Social Psychology*, 20, 484-496.
- Orne, M. T. (1969). Demand characteristics and the concept of quasi-controls. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 143-179). New York: Academic Press.
- Porter, L. W., & Lawler, E. E., III (1968). *Managerial attitudes and performance*. Homewood, IL: Richard D. Irwin.
- Sheppard, B. H., & Lewicki, R. J. (in press). Toward general principles of managerial fairness. *Social Justice Review*.
- Thibaut, J., & Walker, L. (1975). *Procedural justice: A psychological analysis*. Hillsdale, NJ: Erlbaum.
- Thibaut, J., & Walker, L. (1978). A theory of procedure. *California Law Review*, 66, 541-566.
- Tyler, T. R. (1984). The role of perceived injustice in defendants' evaluations of their courtroom experience. *Law and Society Review*, 18, 386-401.
- Tyler, T. R., & Caine, A. (1981). The role of distributional and procedural fairness in the endorsement of formal leaders. *Journal of Personality and Social Psychology*, 41, 642-655.
- Tyler, T. R., & Folger, R. (1980). Distributional and procedural aspects of satisfaction with citizen-police encounters. *Basic and Applied Social Psychology*, 1, 281-292.
- Walker, L., Lind, E. A., & Thibaut, J. (1979). The relation between procedural and distributive justice. *Virginia Law Review*, 65, 1401-1420.
- Walster, E., Walster, G. W., & Berscheid, E. (1978). *Equity: Theory and research*. Boston, MA: Allyn & Bacon.

Received February 3, 1986

Revision received April 15, 1986 ■

Intentionally Favored, Unintentionally Harmed? Impact of Sex-Based Preferential Selection on Self-Perceptions and Self-Evaluations

Madeline E. Heilman, Michael C. Simon, and David P. Repper
New York University

In this laboratory study we compared the effect of sex-based preferential selection with that of merit-based selection on the reactions of 64 male and 76 female undergraduates serving as task leaders. Subjects succeeded or failed on the task while working with another individual (a confederate). As predicted, only women's self-perceptions and self-evaluations were negatively affected by the sex-based preferential selection method relative to the merit-based method. When selected on the basis of sex, women devalued their leadership performance, took less credit for successful outcomes, and reported less interest in persisting as leader; they also characterized themselves as more deficient in general leadership skills. These findings suggest that when individuals have doubts about their competence to perform a job effectively, nonwork-related preferential selection is likely to have adverse consequences on how they view themselves and their performance. Implications of the findings for the implementation of affirmative action programs are discussed.

How people attain leadership roles has been shown to affect how they are viewed by others (Hollander, 1978; Jacobson & Koch, 1977; Read, 1974), and to influence subordinate behavior and performance (Goldman & Fraas, 1965; Walker, 1976). Few investigations, however, have focused on the consequences of different attainment methods on the leaders themselves. It is the objective of this research to examine the consequences of differing leader selection processes on leaders' self-perceptions and self-evaluations. Specifically, we are interested in the impact of leader selection processes preferentially based on sex, as compared with leader-selection processes based on merit.

The question addressed here is interesting not only theoretically, but also because of its broad implications for current personnel practices. In many organizations, affirmative action programs have been developed in response to a legal mandate (Equal Employment Opportunity Commission, 1972, 1979) to engage in an active process of ensuring nondiscriminatory selection and placement procedures. In its most basic form, an affirmative action program involves efforts to expand the applicant pool so that members of minority groups are given an equal opportunity for selection and placement. Often, however, affirmative action programs also include preferential hiring based on federal guidelines or hiring quotas. In these cases, women and other disadvantaged groups are intentionally favored when selection and placement decisions are made. Although the objective of such hiring programs is to remedy the ills of past discriminatory practices, some have argued that they harm the very people they were ostensibly designed to benefit

(Sowell, 1978). There are data that lend support to this view. Heilman and Herlihy (1984) have shown that women's jobs are devalued when they are thought to have obtained them through preferential selection, and Jacobson and Koch (1977) have demonstrated the negative consequences of sex-based preferential selection when a leaders' performance is judged by subordinates. In addition, Chacko (1982) has presented correlational data linking women managers' perceptions that they were hired because of their sex, with low organizational commitment and job satisfaction and with high role stress. But perhaps most central to the question of whether such selection procedures harm those they are designed to benefit is the issue that is the focus of this investigation: the effects of preferential selection on the self-perceptions and self-evaluations of those who are intentionally favored.

Why might preferential selection have different consequences for how one views oneself than selection based on merit? To answer this question, it is necessary to examine the message these differing selection procedures convey to the individual about his or her job competence. Preferential selection implies that a work-irrelevant characteristic had special weight in the decision process, whereas merit-based selection implies that skill and ability were the critical deciding points. Consequently, those selected on the basis of merit have received a vote of confidence from the powers that be—they feel they have earned their positions and their sense of competence is affirmed. Those selected on the basis of preference, however, have not received such external verification of capability; instead, their competence is open to question. For this reason, those in preferential selection situations are vulnerable to feelings of inadequacy for the job in which they are placed.

But, not having explicit verification about qualifications for a job need not always result in negative feelings about one's capacity to handle it effectively. If an individual feels competent to do the job to begin with, then verification of his or her skills and abilities is unnecessary; unless directly challenged, one's

This research was supported by Grant GA EO 8521 from the Rockefeller Foundation.

The authors wish to thank Jonathan Lucas and Laurie Miller for their many important contributions to this study.

Correspondence concerning this article should be addressed to Madeline E. Heilman, Room 550, Department of Psychology, New York University, 6 Washington Place, New York, New York 10003.

perceptions of competence will persist. If, however, an individual harbors self-doubts and negative performance expectations, the absence of such verification is apt to fuel these perceptions further. Therefore, the ambiguity about competence established by preferential selection procedures is apt to more adversely affect those who lack confidence in their ability to perform the job well.

Based on this reasoning, one might expect differences in the way people react to preferential selection as compared with merit-based selection to leadership positions. Not only individual differences but also group membership may determine these reactions. Consider, for instance, women and men. In our culture, leadership positions typically are considered to be male in character (Schein, 1973, 1975). And not surprisingly, societal stereotypes about women and men depict women as having far fewer of the qualities that comprise effective leadership skills than do men (e.g., Broverman, Vogel, Broverman, Clarkson, & Rosenkrantz, 1972). These stereotypes are widely shared, and evidence exists that both men and women accept them as self-defining (e.g., Rosenkrantz, Vogel, Bee, Broverman, & Broverman, 1968). Consequently, men, as a group, are likely to approach leadership roles with confidence in their ability and thus high expectations of success. In contrast, women, as a group, are likely to approach leadership roles with a lack of self-confidence and thus low expectations of success. Support for this point can be found in the work of Lenny (1977) and McMahan (1982), among others. Given this, one would expect that whether leader selection was based on merit or nonmerit factors should have little impact on men but a great deal of impact on women: Whereas merit-based selection should allay women's concerns about their competence, preferential-selection procedures should exacerbate them.

If the reason preferential selection has negative consequences for women assigned to leadership positions is their underlying lack of confidence in their leadership abilities, then this negativity might well result no matter what the basis of preferential treatment. However, the negative impact should most certainly be evident when sex is its basis. Under these conditions, stereotypes about women are apt to be salient, and the perceived gap between the attributes women bring to a leadership role and those required to be an effective leader is likely to be great. Thus, because it focuses on the very reason for their lack of confidence, sex-based preferential treatment in selection decisions should be particularly devastating for women, deepening their self-doubts and giving rise to pessimism about their future performance.

It is interesting to consider whether actual performance overrides the unfavorable self-view of competence that can arise from preferential leader selection. If the leader is ultimately successful, will the negativity provoked by the selection procedure dissipate? If so, then the problems presented thus far are somewhat less grave than they appear at this point. Unfortunately, however, given what we know about performance expectations and the power they exert on the interpretation of subsequent performance data, this seems highly unlikely. Generally, we tend to perceive the world in a manner that proves our performance expectations correct, whether they are positive or negative (Diggory, 1966; Feather, 1966). This tendency should be particularly evident in leadership situations because the behav-

iors constituting effective work are often ambiguous, task accomplishment frequently requires the efforts of other than oneself, and the cause of the outcome can easily be cognitively adjusted to remain consistent with the self-view underlying the performance expectations. Thus, it is likely that despite general differences in reactions to success and failure outcomes, the differential effects of the two leader-selection procedures on the sense of competence of female leaders will be evident regardless of whether their performance is actually successful or unsuccessful.

Research leads us to believe that differences in self-view of competence and the performance expectations they generate will influence how an individual rates his or her own performance effectiveness and his or her causal attributions for success and failure (Deaux, 1976). In addition, it has been suggested that such self-views and expectations affect aspirations and striving (Weiner et al., 1972); consequently, leaders who see themselves as incompetent and expect to do poorly should be more likely to seek to remove themselves from the position than those who see themselves as competent and expect to do well. In light of this, it is predicted that when selected on the basis of sex rather than merit, women leaders will not only regard themselves as less capable, but also will derogate their performance and be more willing to relinquish their leadership roles. This would be expected regardless of performance outcome. However, the extent to which personal responsibility is taken should differ as a function of outcome. Consistent with the idea that sex-based leader selection results in a negative self-view for women leaders, these women should be less accepting of credit for success and more accepting of blame for failure than those selected on the basis of merit.

In the following study, male and female subjects were assigned to leadership roles as a result of either their skills and abilities or their sex. They performed a one-way communication task in which they took an active leadership role and a confederate served as a follower. After completing the task, feedback about the task outcome as successful or unsuccessful was provided. It was expected that the different leader selection processes would have consequences for women, but not for men; furthermore, these consequences were expected whatever the task outcome. Women in preferential selection conditions as compared with those in merit-based selection conditions were expected (a) to judge their general leadership abilities less favorably, (b) to evaluate their performance less favorably, (c) to take less responsibility for successful performance outcomes and more responsibility for unsuccessful ones, and (d) to be less desirous of continuing in the leadership position. Reactions to the task and to the selection procedure, and self-reports of stress and motivation, also were obtained to determine the breadth of the consequences of the leader selection variable.

Method

Subjects and Design

Subjects were 140 (76 female and 64 male) undergraduates who participated in the research for partial fulfillment of an introductory psychology course requirement at New York University. They ranged from 18 to 22 years of age. The design was a $2 \times 2 \times 2$ factorial involving three independent variables: sex of subject (male or female), assignment

method (merit-based or sex-based preferential selection) and task outcome (success or failure). Nineteen female and 16 male subjects were randomly assigned to each of the four experimental conditions. The different numbers of men and women participating in the study was a consequence of the disproportionate numbers of men and women enrolled in the course.

Procedure

Subjects were met by one of two male experimenters and were informed that the research would begin as soon as the other participant arrived. When the other participant (actually a confederate posing as another student) arrived, the experimenter escorted them into the research room where they were guided to seats at desks on different sides of the room; they sat with their backs to one another. The confederate was always of the sex opposite to that of the subject.

The experimenter introduced the study as part of a research project concerning leadership and communication. Subjects were told that they would be either a leader or a follower in a one-way communication task in which the leader would instruct the follower in the drawing of geometric figures. It was stated that the leader would have the more creative and visible position, although task performance ultimately depended on both leader and follower. Subjects were then asked to complete the 14-item Spatial Communication Skills Inventory (SCSI), which was said to reliably assess their one-way communication abilities. The questions in the SCSI concerned both spatial relations and communication skills, and were constructed for purposes of this research.

After completing the SCSI, the one-way communication task was explained. (For some subjects the SCSI was apparently scored by the experimenter; for others, it was not. See the assignment method manipulation in the next section.) Subjects were told that whoever was selected as the leader would verbally instruct the follower in the drawing of each of three complex geometric figures. In order to perform the task, it was explained, the leader would be given 10 s to decide on an appropriate strategy and then would be given 2 min to direct the follower in the drawing of the figure. Because the task was to involve one-way communication, subjects were to remain with their backs to one another and followers were not to respond to the leader's instructions nor ask questions of clarification; only the leader was to talk during the task period. The necessity of the leader's precision and general skill in one-way communication was underscored, as was the follower's attentiveness.

The figures, which were developed expressly for this task, were complex designs composed of multiple lines, curves, and angles. It was established in pilot work that none of the figures could be completed in the 2-min task period. Subjects were told that the drafted figures would be scored on the basis of both accuracy and completeness, and the process by which the scores were to be determined was explained.

At this point, the subject and confederate were assigned to roles and the assignment method manipulation was enacted. The subject always was made the leader, and the confederate, the follower. The task materials were then distributed and the task began. On completion of the task, the experimenter collected the drafted figures. The experimenter asked the subject and confederate to refrain from talking to one another for the few moments it would take to score them. He then spent some time at his desk, apparently evaluating the figures.

After the experimenter ostensibly had finished scoring the figures, subjects received information about their task outcome and, after a moment to absorb it, were given two brief questionnaires to complete containing both manipulation checks and dependent variable measures. The first concerned their role in the experiment and the second, their impressions of the one-way communication task and the research session more generally.

After collecting the completed questionnaires the experimenter explained the experimental manipulations and the purpose of the research, and answered any questions the subjects might have.

Experimental Manipulations

Assignment method. After completing the SCSI, subjects in merit conditions sat for a moment while the experimenter supposedly scored the inventories. The inventories of those in preferential treatment conditions were simply collected and conspicuously set aside by the experimenter.

After the task instructions were completed, the leader and follower roles were assigned. In all conditions, subjects were told the following:

Normally, in situations like this, leaders are selected on the basis of skill and ability, which basically means that they are good at the task. We've been doing our selecting this way also, by using the Spatial Communication Skills Inventory that you just finished. It is a highly reliable measure of one-way communication skills developed by psychologists.

What followed differed in the merit and preference conditions. In merit conditions, the experimenter indicated that the subject had performed well and had scored better than the confederate. He then said,

So, *you* [pointing at the subject], since you scored better on the inventory, will get to be leader for the task.

However, in preference conditions, the experimenter said,

But today we are going to have to do things a little differently, because there just haven't been enough male (female) subjects signing up so far. So, regardless of how each of you did on the inventory, *you* [pointing at the subject], since you're a man (woman), will get to be leader for this task.

To reinforce the desirability of the leadership role, the confederate audibly sighed when told that he or she would be the follower.

Outcome. The outcome manipulation was accomplished through the verbal feedback given about the drafted figures. In success and failure conditions, subjects were told, respectively,

It seems that the two of you have done pretty well (haven't done very well). Your score was a 36 for accuracy and completeness combined . . . You definitely fall into our top (bottom) quartile, which means that you fall in with the top (bottom) 25% of all twosomes who have done this task.

Dependent Measures

To determine how subjects perceived their leadership ability, they were asked how they generally perform on tasks involving leadership skills. Responses were made on a 9-point scale, the endpoints of which were *very well* and *very poorly*. A 9-point scale also was used to ascertain subjects' desire to persist in the leadership position: They were asked to indicate how much they would want to be leaders in a subsequent task session, and responded on a scale ranging from *very much* to *not at all*. Subjects were also asked to divide up the total responsibility for the performance outcome (100%) between themselves and the other participant.

The performance evaluation measure was a composite measure based on a number of individual item scales. First, four 9-point bipolar adjective scales, which were selected a priori to measure self-perceptions of performance competence on the leadership task (competent-incompetent, effective-ineffective, strong-weak, and decisive-indecisive), were combined into one index by averaging the four ratings (coefficient $\alpha = .86$). Second, one item focused on performance directly: Subjects were asked to rate how they performed as a leader on a 9-point scale, the endpoints of which were *very well* and *very poorly*. Because the correlation between the performance competence measure and the performance rating was quite high ($r = .76$), and separate analyses of the two

Table 1
Intercorrelations Between the Dependent Measures

Measure	1	2	3	4	5	6	7	8
1. Leadership ability	—	.33	.02	.30	.13	.28	.08	.16
2. Performance evaluation		—	-.27	.52	.21	.28	.22	.23
3. % responsibility			—	-.14	-.02	-.07	-.03	.01
4. Desire to remain leader				—	.10	.21	.18	.38
5. Involvement level					—	-.02	.68	.12
6. Stress level						—	.06	.10
7. Task reaction							—	.14
8. Perceived fairness								—

measures yielded identical results, a combined variable was created by averaging across them (coefficient $\alpha = .85$). This composite variable was used as our measure of performance evaluation.

Subjects also reported how they felt when working on the task. A stress index was derived from self-ratings on three scales: calm–nervous, stressed–not stressed, and relaxed–tense (coefficient $\alpha = .87$). An involvement index was derived from self-ratings on three other scales: motivated–unmotivated, uninvolved–involved, and interested–bored (coefficient $\alpha = .74$). An overall task reaction measure was derived from subjects' responses to four bipolar adjective scales: pleasant–unpleasant, enjoyable–unenjoyable, interesting–boring, and dull–stimulating (coefficient $\alpha = .85$). Subjects also were asked to rate the fairness of the selection procedure. In all cases, 9-point scales were used in obtaining these data.

The intercorrelations among all the dependent measures appear in Table 1.

Results

Manipulation Checks

When asked how they had been rated on the task, virtually all (99.3%) of the subjects correctly indicated the quartile representing their manipulated performance outcome. The one subject who wrongly indicated the manipulated outcome was in a success condition. The manipulation of the assignment variable also appeared to have its intended effect. In response to a question about how the leader was selected, all but 2 of the subjects indicated the correct method of assignment (assignment on the basis of a test of ability or skill, or on the basis of factors other than ability or skill). Of the 2 subjects who indicated a leader assignment method contrary to condition, one was in a preferential treatment condition and the other was in a merit condition.

Dependent Measures

A multivariate analysis of variance was conducted on the four scale ratings composing the key dependent measures (perception of leadership ability, performance evaluation, assignment of responsibility, and desire to persist in the leadership role). Overall, the multivariate F was significant for outcome, $F(4, 129) = 17.57, p < .0001$; assignment method, $F(4, 129) = 5.62, p < .0003$; sex of subject, $F(4, 129) = 2.91, p < .03$; and most important for our hypotheses, the interaction between assignment method and sex of subject, $F(4, 129) = 2.76, p < .03$. Having established the overall effects, univariate analyses of

variance (ANOVAS) were conducted and, when appropriate, were followed up by simple effects tests for purposes of clarification. Because preliminary analyses indicated no significant differences in subjects' responses as a function of who the experimenter was, data from sessions run by both experimenters were combined for all of the analyses. The means for each of the four dependent variables are presented in Table 2.

Perception of leadership ability. Analysis of variance of subjects' ratings of their general leadership ability indicated a main effect for assignment method, $F(1, 132) = 4.64, p < .04, n^2 = .032$, and an Assignment Method \times Sex of Subject interaction, $F(1, 132) = 5.66, p < .02, n^2 = .039$. Simple effects tests made clear the source of this interaction: The method of assignment had effects only on the ratings of women, $F(1, 132) = 10.21, p < .01, n^2 = .07$, not on the ratings of men, $F(1, 132) = 0.09, ns$. As was predicted, preferential treatment had a deleterious effect on female subjects' perceptions of their leadership ability (whatever the task outcome), but had little effect on the self-perceptions of the men.

Performance evaluation. Results of the ANOVA of the performance evaluation composite indicated a main effect for outcome, $F(1, 132) = 52.92, p < .0001, n^2 = .257$, with, as might be expected, successful subjects rating their performance more favorably than failing ones. In addition, and of more interest to the aims of this investigation, the results revealed significant main effects for assignment method, $F(1, 132) = 4.42, p < .04, n^2 = .022$; sex of subject, $F(1, 132) = 5.79, p < .02, n^2 = .028$; and an Assignment Method \times Sex of Subject interaction, $F(1, 132) = 7.59, p < .01, n^2 = .037$. Simple effects tests demonstrated that, as predicted, preferential selection resulted in less favorable performance evaluations than did merit-based selection only for female subjects, $F(1, 132) = 11.63, p < .001, n^2 = .057$, not for male subjects, $F(1, 132) = 0.37, ns$. Thus, once again, the negative effects of preferential selection were confined to women, and were not dependent on task outcome.

Responsibility for the task outcome. Analysis of the percentages of the total responsibility for the task outcome claimed by subjects revealed a main effect for outcome, $F(1, 132) = 24.91, p < .0001, n^2 = .141$, indicating that subjects generally took more responsibility on themselves for failure than for success. Moreover, an assignment method main effect, $F(1, 132) = 12.70, p < .0005, n^2 = .072$, indicated that those in preferential selection conditions took less responsibility than those in merit-based selection conditions. However, a significant three-way in-

Table 2
Means of Key Dependent Variables in Each Experimental Condition

Variable	Leadership ability ^a	Performance evaluation ^a	% responsibility assigned to self	Desire to remain leader ^a
Male success				
Merit	6.44	6.22	57.19	6.94
Preference	6.75	6.06	54.12	6.25
Female success				
Merit	6.79	6.34	60.00	6.47
Preference	5.53	4.79	45.79	4.68
Male failure				
Merit	6.50	4.08	71.25	4.75
Preference	6.44	4.68	59.69	5.31
Female failure				
Merit	6.63	4.13	66.05	4.53
Preference	5.42	3.24	63.68	3.32

^a The higher the mean, the more favorable the rating.

teraction, $F(1, 132) = 5.30, p < .03, n^2 = .03$, and an inspection of the means in Table 2 prompted simple effects tests to clarify the effects of assignment method. Results indicated that the simple main effect for assignment method held only among women in success conditions, $F(1, 132) = 11.31, p < .001, n^2 = .064$, and men in failure conditions, $F(1, 132) = 6.31, p < .025, n^2 = .036$; neither the simple main effect for women in failure conditions, $F(1, 132) = .036, ns$, nor for men in success conditions, $F(1, 132) = 0.44, ns$, were statistically significant. These results are nonsupportive of our predictions in two respects: First, preferential treatment did not have negative consequences for all female subjects regardless of task outcome, and second, some of the male subjects were unexpectedly found to differ in their responses as a result of the method by which they were assigned to the leadership role.

Desire to persist in the leadership position. Analysis of variance of subjects' reported desire to remain in the leadership position revealed a main effect for outcome, $F(1, 132) = 17.65, p < .0001, n^2 = .106$, with those in success conditions expressing more desire to persist as leaders than those in failure conditions. In addition, and of more importance to the ideas under investigation, the results indicated a main effect for assignment method, $F(1, 132) = 4.81, p < .03, n^2 = .029$; a main effect for sex of subject, $F(1, 132) = 7.59, p < .01, n^2 = .045$; and a marginally significant interaction between them, $F(1, 132) = 3.47, p < .07, n^2 = .021$. Simple effects tests were conducted to further explore these results, and revealed that the main effect for assignment method held only for female subjects, $F(1, 132) = 8.25, p < .01, n^2 = .09$, not for male subjects, $F(1, 132) = 0.01, ns$. Thus, as we had predicted, preferential treatment had a consistently negative effect on the desire to persist as leader when subjects were women, but not when they were men.

Additional Findings

Several additional measures were obtained to aid in interpreting responses to the key dependent variables and, also, to provide information about the scope of the effects of the method of leader assignment on the individual. Means of responses to these measures appear in Table 3.

Feelings while working on the task. Analysis of variance of the involvement composite scores revealed a main effect for outcome, $F(1, 132) = 17.07, p < .001, n^2 = .105$, and a three-way interaction effect, $F(1, 132) = 5.64, p < .02, n^2 = .035$. Examination of the means and simple effects tests make clear that outcome only made a difference in subjects' reports of their involvement when subjects were men in merit conditions, $F(1, 132) = 23.31, p < .001, n^2 = .144$. In particular, it is the male subjects in the merit failure conditions whose reports appear idiosyncratic. As Table 3 makes evident, and subsequent post hoc Tukey contrasts verify, the mean in this condition is markedly lower than the means in all of the other conditions (all comparisons, $p < .05$).

Analysis of variance of the stress composite scores indicated no significant effects whatsoever.

Reactions to the task. Reactions to the task were generally very favorable ($M = 6.5$). Results of the ANOVA of the task reaction composite scores indicated no effects other than outcome, $F(1, 132) = 20.88, p < .0001, n^2 = .127$, with those in success conditions rating it more favorably than those in failure conditions.

Perceived fairness of the leader selection method. Analysis of variance of subjects' fairness ratings revealed a main effect for assignment method, $F(1, 132) = 41.04, p < .0001, n^2 = .225$, indicating that those assigned on the basis of merit viewed the assignment method as more equitable than those appointed on the basis of preference. Also indicated was an outcome effect, $F(1, 132) = 4.92, p < .03, n^2 = .027$, indicating that those in failure conditions rated the selection procedure as less fair than those in success conditions. However, a marginally significant interaction effect, $F(1, 132) = 3.18, p < .08$, and follow-up simple effects tests, made clear that this "sour grapes" phenomenon occurred only in merit conditions, $F(1, 132) = 8.02, p < .01, n^2 = .044$; in preferential treatment conditions, no difference was evident between those subjects who believed they had failed at the task and those who believed they had succeeded, $F(1, 132) = .09, ns$.

Discussion

These results provide strong evidence that sex-based preferential selection procedures can have detrimental effects on lead-

Table 3
Means of Self-Perceived Involvement, Stress Index, Task Reaction, and Fairness

Variable	Involvement level	Stress index	Task reaction	Perceived fairness
Male success				
Merit	7.67	4.79	7.48	7.00
Preference	7.04	5.31	6.67	4.44
Female success				
Merit	7.26	5.49	6.74	7.11
Preference	7.02	4.86	7.22	4.16
Male failure				
Merit	5.40	4.90	6.41	5.62
Preference	6.69	5.19	6.19	4.25
Female failure				
Merit	6.82	5.00	6.09	5.79
Preference	6.35	4.39	6.34	4.05

Note. The higher the mean, the more favorable the rating.

ers' self-perceptions and self-evaluations, and that these detrimental effects occur despite information about performance quality. When selected on the basis of preference rather than merit, the women consistently rated their performance more negatively, took less credit for successful outcomes, and were less eager to persist in their leadership roles; they also viewed themselves as more deficient in general leadership skills. However, these results also clearly demonstrate that sex-based preferential selection does not have detrimental consequences for everyone. As predicted, the method of leader selection had effects only on women, not on men.

It had been proposed that sex-based preferential selection procedures would have negative consequences only for women because they, as a group, are unlikely to be confident about their ability to succeed in a leadership position, whereas men, as a group, are confident about their ability in this regard. Women's performance apprehensions were therefore expected to be exacerbated by the ambiguity of preferential selection and alleviated by the reaffirmation of competence inherent in merit-based selection, giving rise to differential evaluations of self and performance. Our results are consistent with this idea and, indeed, lend support to it.

Note that our results demonstrated that sex-based preferential selection not only affected female subjects' reactions related to the task at hand, but also affected their sense of general leadership ability. This is a very important finding. It suggests that the negative self-perceptions induced by sex-based preferential processes are not confined to a particular performance instance but are more general in character. This idea is further supported by the strong indication that women selected preferentially desired to remove themselves from the leadership position, suggesting that motivation to remain in a leadership role also was negatively affected. Clearly the harm that can result from preferential selection processes is not negligible in magnitude nor in scope.

Subjects' indications of their personal responsibility for the performance outcome contained some unexpected findings. First, although women selected preferentially took less credit for success, they did not take more blame for failure than did those selected on the basis of merit. This is curious because, regardless of performance outcome, these women clearly derogated

both their performance and their ability, and personal responsibility ratings should reflect these evaluations. One possible explanation for this apparent anomaly lies in our subjects' atypical tendency to take responsibility for failure. Whereas failure usually is attributed externally, some aspect of our task or research situation seemed to prompt internal attributions, and in all instances subjects took more than 50% of the blame for failure. This overall willingness to accept responsibility for failure may have obscured the differences we had predicted. A second unexpected finding marked the single instance in which male subjects differed in their responses as a function of the leader selection method: They took more blame for failure when selected on the basis of merit than when selected preferentially. Because we expected the responsibility data to reflect self-evaluations of performance and ability, this finding was counter to our expectations. However, an examination of the self-reports of involvement while working on the task provides some insight. Although men selected on the basis of merit did, in fact, take a good deal of responsibility for failure, they also reported themselves to have been less involved than anyone else while working on the task. Thus, these men viewed the failure as their doing, but attributed it to motivational, not ability, deficits.

The lack of differences in task reactions and self-reports of stress while working suggest that the negative effects of sex-based preferential selection do not extend to emotional experience or work attitudes. However, caution is urged in making such an extrapolation from these data. Our task, although a laboratory one, proved to be extremely engaging and high in impact. Furthermore, reactions to it were very favorable. It is conceivable that with a less engaging or more unpleasant task, additional effects of the leader selection procedure might become apparent. As it is, the discrepancy in fairness attributed to the two selection procedures suggests other types of work-related attitudes that might be affected by preferential treatment processes.

Although in this study it was the women who were expected to lack confidence in their abilities and therefore be most vulnerable to the negative effects of sex-based preferential selection, under the right conditions men should be similarly affected. What is critical is not the sex of the individual but the degree to which he or she is confident of his or her ability to

perform a job well. Thus, if a task were a traditionally feminine one, men who are preferentially selected on the basis of sex should undergo the same problems as did the women in our study. Furthermore, we do not wish to imply that all women and all men differ from one another with respect to confidence, and that these differences are induced solely by elements of the situation. On the contrary, individual differences in self-confidence are apt to be very powerful determinants of whether preferential selection will have adverse consequences for the individuals involved. Finally, although we have focused on sex-based preferential selection in this investigation, the work reported here is of relevance to any type of nonwork-based preferential selection. The effects of racially based preferential selection or of nepotism, for instance, should be regulated by the same dynamics as those outlined here. Whenever individuals harbor doubts about their competence, regardless of whether such doubts are warranted, preferential selection on the basis of nonwork-related criteria is likely to have deleterious consequences for their self-perceptions and self-evaluations.

This research has important implications for current personnel policies and practices. Although it is rare that competence is not at all a factor when affirmative action plans are in operation, the common perception is that it is not a major criterion for selection. Our results suggest that when this is the case, sex-based preferential selection can trigger a vicious cycle of negative self-regard for women targeted for favored treatment. Ironically, this may happen to a woman even when she truly is qualified for the position and would have been hired had merit been the sole basis of selection. However, the findings presented here by no means imply that such adverse consequences inevitably result from affirmative action programs. Rather, they underscore the necessity of paying heed to the way in which such programs are implemented and, in particular, of making sure that selection without regard to competence is not believed to characterize affirmative action efforts. This suggests not only that competence considerations should be a dominant factor in selection decision making, but also that selectees should be made aware of the important role competence played in their selection. For the message in the data seems clear indeed: If affirmative action is associated with an absence of quality standards, its intended beneficiaries may in fact become its victims.

References

- Broverman, I. K., Vogel, S. R., Broverman, D. M., Clarkson, F. E., & Rosenkrantz, P. S. (1972). Sex-role stereotypes: A current reappraisal. *Journal of Social Issues*, 28, 59-78.
- Chacko, T. I. (1982). Women and equal employment opportunity: Some unintended effects. *Journal of Applied Psychology*, 67, 119-123.
- Deaux, K. (1976). Sex: A perspective on the attribution process. In J. H. Harvey, W. Ickes, & R. F. Kidd (Eds.), *New directions in attribution research* (Vol. 1, pp. 335-352). Hillsdale, NJ: Erlbaum.
- Diggory, J. (1966). *Self-evaluation: Concepts and studies*. New York: Wiley.
- Equal Employment Opportunity Commission (1972). Guidelines on discrimination because of sex. *Federal Register*, 37, 6836.
- Equal Employment Opportunity Commission (1979, January 19). *Affirmative action guidelines* (Report No. 44 FR 4421). Washington, DC: Author.
- Feather, N. T. (1966). Effects of prior success and failure on expectations of success and subsequent performance. *Journal of Personality and Social Psychology*, 3, 287-298.
- Goldman, M., & Fraas, L. A. (1965). The effects of leader selection on group performance. *Sociometry*, 28, 82-88.
- Heilman, M. E., & Herlihy, J. M. (1984). Affirmative action, negative reaction? Some moderating conditions. *Organizational Behavior and Human Performance*, 33, 204-213.
- Hollander, E. P. (1978). *Leadership dynamics*. New York: Free Press.
- Jacobson, M. B., & Koch, W. (1977). Women as leaders: Performance evaluation as a function of method of leader selection. *Organizational Behavior and Human Performance*, 20, 149-157.
- Lenny, E. (1977). Women's self-confidence in achievement settings. *Psychological Bulletin*, 84, 1-13.
- McMahan, I. D. (1982). Expectancy of success on sex-linked tasks. *Sex Roles*, 8, 949-958.
- Read, P. B. (1974). Source of authority and legitimization of leadership in small groups. *Sociometry*, 37, 189-204.
- Rosenkrantz, P. S., Vogel, S. R., Bee, H., Broverman, I. K., & Broverman, D. M. (1968). Sex-role stereotypes and self-concepts in college students. *Journal of Consulting and Clinical Psychology*, 32, 287-295.
- Schein, V. E. (1973). The relationship between sex role stereotypes and requisite management characteristics. *Journal of Applied Psychology*, 57, 95-100.
- Schein, V. E. (1975). Relationships between sex role stereotypes and requisite management characteristics among female managers. *Journal of Applied Psychology*, 60, 340-344.
- Sowell, T. (1978, June). Are quotas good for blacks? *Commentary*, 65, 39-43.
- Walker, T. G. (1976). Leader selection and behavior in small political groups. *Small Group Behavior*, 7, 363-368.
- Weiner, B., Frieze, I. H., Kukla, A., Reed, L., Rest, S., & Rosenbaum, R. M. (1972). Perceiving the causes of success and failure. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 95-120). Morristown, NJ: General Learning Press.

Received November 25, 1985

Revision received May 16, 1986 ■

A Revision of the Job Diagnostic Survey: Elimination of a Measurement Artifact

Jacqueline R. Idaszak and Fritz Drasgow
University of Illinois, Urbana-Champaign

The dimensionality of the original Job Diagnostic Survey (JDS) and a revision were investigated. Factor analyses of two data sets identified six dimensions underlying the original JDS. Five of the factors correspond to the pattern expected for the JDS items; the sixth was identified as a measurement artifact. Five of the JDS items were subsequently rewritten to eliminate the artifact. The revised survey was administered to employees of a printing company ($N = 134$) and the *a priori* five-factor solution was obtained with no artifact factor. Scale-factor correlations were also computed. The resulting coefficients suggest that the revised JDS scales are measuring their underlying constructs with reasonable accuracy. As a result of the measurement artifact in the original JDS, it is recommended that the revised JDS should be used in future research concerned with task characteristics.

Characteristics of jobs play a central role in organizational theory. They can be viewed as "technology's most direct consequences" (Hulin & Roznowski, 1985, p. 71). Job enlargement and job enrichment programs typically treat job characteristics as the independent variables that should be manipulated. More generally, a variety of organizational theories hypothesize that job characteristics are precursors of job-related affect, productivity, and withdrawal (Hackman & Oldham, 1974, 1975; Mowday, Porter, & Steers, 1982; Turner & Lawrence, 1965).

At present, the most popular perceptual measure of job characteristics seems to be Hackman and Oldham's (1974, 1975) Job Diagnostic Survey (JDS). Its popularity, however, is more a consequence of Hackman and Oldham's theory of job characteristics (upon which the JDS is based) than the psychometric properties of the instrument itself. Specifically, several questions remain unanswered with respect to the latent structure of the JDS. Until recently, evidence that the JDS measures the hypothesized dimensions was weak (Dunham, 1976; Dunham, Aldag, & Brief, 1977; Fried & Ferris, 1986; Green, Armenakis, Marbert, & Bedeian, 1979; Pierce & Dunham, 1978; Pokorney, Gilmore, & Beehr, 1980).

Dunham and his colleagues (Dunham, 1976; Dunham et al., 1977; Pierce & Dunham, 1978) and Pokorney et al. (1980) looked at the factor solutions for a wide variety of samples and found very few that resembled the *a priori* five-factor structure. Dunham (1976), for example, advocated a single-factor solution representing job complexity. Other studies have accepted two-, three-, and four-factor solutions in addition to the rarely encountered five-factor structures (Dunham et al., 1977; Fried

& Ferris, 1986; Green et al., 1979; Pierce & Dunham, 1978; Pokorney et al., 1980). From this set of studies it would seem appropriate to conclude that the JDS should be empirically examined in each new subpopulation.

In an attempt to explain these inconsistencies, Fried and Ferris (1986) investigated possible moderators of the underlying JDS factor structure. Based on their results they suggested that age, education level, and position level influence the factor structure.

Harvey, Billings, and Nilan (1985) used a different approach in an attempt to resolve the JDS dimensionality issue. They used confirmatory factor analysis to evaluate the factor structures suggested in past research. Their results suggest that none of the factor structures found in the past provide an adequate fit. Instead, an oblique solution with Hackman and Oldham's *a priori* dimensions plus one or two method factors (factors for the reverse-scored items and the three-anchor scale items) provided the best fit. As a result of the existence of these measurement artifacts, it was concluded that the JDS, in its present state, is psychometrically "troublesome." Furthermore, a revised version of the JDS should be developed that eliminates the measurement artifact.

Although Harvey et al. (1985) provide a plausible explanation for the inconsistencies found in past JDS research, it is still possible, as Dunham et al. (1977) suggested, that the results of JDS factor analyses are sample specific. Further evidence from heterogeneous samples is needed to verify the measurement artifacts suggested by Harvey et al.

The research described in this article provides replications of Harvey et al.'s (1985) findings with two separate samples. In Study 1, the responses of a heterogeneous subsample from Oldham, Hackman, and Stepina's (1978) JDS data base were selected and factor analyzed. Study 2 cross validated the factor structure identified in Study 1, using an independent subsample. In Study 3, several JDS items were revised in an attempt to correct the measurement artifact identified in Studies 1 and 2. The revised items were administered to a new sample, and the data set was factor analyzed. We obtained a five-factor solution that closely reflected Hackman and Oldham's theory.

We extend special thanks to Charles L. Hulin for his helpful comments on earlier versions of this article and to Greg R. Oldham and J. Richard Hackman for their help in revising the Job Diagnostic Survey items.

Correspondence concerning this article should be addressed to Jacqueline R. Idaszak, Department of Psychology, University of Illinois, Psychology Building, 603 East Daniel Street, Champaign, Illinois 61820.

Table 1
Six-Factor Maximum Likelihood Factor Structure Using the First Sample From the 1978 JDS Data Base

Item	Factor					
	1	2	3	4	5	6
Autonomy						
1	25 (04)			−08 (04)	58 (04)	
2	18 (05)		44 ^a (03)		31 (04)	
3					72 (03)	
Task Identity						
4	32 (03)	50 (03)		−11 (04)		
5		78 (06)	39 ^a (04)		−22 (06)	−19 (05)
6		70 (03)				
Skill Variety						
7	74 (02)					
8	67 (02)					
9	63 (05)		37 ^a (03)		−24 (04)	
Task Significance						
10	23 (04)	08 (03)		47 (04)		
11				66 (04)		
12			42 ^a (03)	40 (04)		
Feedback						
13				−13 (04)		83 (04)
14						73 (03)
15	−14 (05)		44 ^a (04)			46 (04)
Factor correlation matrix						
1	—					
2	22 (04)	—				
3	30 (08)	−05 (05)	—			
4	49 (04)	19 (05)	04 (07)	—		
5	55 (04)	50 (04)	21 (07)	31 (05)	—	
6	49 (03)	41 (04)	23 (06)	53 (04)	50 (03)	—

Note. All factor loadings fixed at zero are omitted; decimal points are omitted; standard errors of parameter estimates are in parentheses. JDS = Job Diagnostic Survey.
^a This item is a reverse-scored item on the original JDS.

General Analysis

For each study the following analyses were conducted. First, a reduced correlation matrix, using squared multiple correlations as communality estimates, was obtained for the 15 JDS items. Principal axes factor analyses were computed for one through six dimensions. A parallel analysis (Humphreys & Montanelli, 1975) was performed on the eigenvalues. Then each principal axes solution was rotated orthogonally by varimax and obliquely by the direct artificial personal probabilities factor rotation (DAPPPR) of Tucker and Finkbeiner (1981). In the next step of the analysis, the DAPPPR solutions were used as starting values for maximum likelihood factor analyses computed by the LISREL IV computer program. A minimally identified solution was first obtained. Then small loadings were fixed at zero according to the Kroonenberg and Lewis (1982) procedure. In this procedure, *t* statistics provided for the loadings in the LISREL IV output were examined and all loadings associated with small *t* values were set to zero. Chi-square statistics were used as measures of the goodness of fit of the various solutions. The first derivatives of loadings fixed at zero were also used as measures of fit. A ratio criterion of increase in chi-square to increase in degrees of freedom was used to select the “best” restricted solution. Analyses were stopped when a large ratio was obtained.

The final analysis performed provided a measure of the extent to which scale scores correspond to the underlying constructs. The correlations between scales and factors, termed *fidelity coefficients* (Drasgow & Miller, 1982), were computed for each scale derived from the principal axes factor analysis solutions and the restricted maximum likelihood factor analysis solutions.

Study 1

Purpose

The first study was performed to investigate whether measurement artifacts distort the factor structure of the JDS.

Method

Subjects. Data for Study 1 consisted of subsamples from the Oldham et al. (1978) JDS data base randomly selected within job category. This sample (*N* = 1,672) consisted of 377 professionals, 380 managers, 416 clerical workers, 336 processing workers, and 163 machine trade workers. These categories were chosen because of their large sample sizes and the heterogeneity of the group. Of the subjects, 53% were women; 60% were between the ages of 20 and 40 years; 49% had a high school degree or less education; and 47% had some technical or college education or a degree from a technical school or a college.

Table 2
JDS Scale–Factor Correlations From the Restricted
Maximum Likelihood Factor Analysis

Scale	Factor					
	1	2	3	4	5	6
Study 1						
Task Identity	.80	.10	.25	.29	.37	.14
Task Significance	.17	.75	.50	.47	.33	.26
Skill Variety	.13	.39	.85	.40	.40	.41
Feedback	.31	.37	.38	.83	.41	.37
Autonomy	.35	.27	.58	.44	.79	.38
Reverse scored	.31	.32	.54	.47	.40	.71
Study 3						
Task Identity	.89	–.11	.21	.23	.37	
Task Significance	.08	.87	.16	.33	.15	
Skill Variety	.25	.19	.82	.44	.37	
Feedback	.26	.23	.24	.86	.42	
Autonomy	.35	.13	.34	.39	.80	
Revised items	.53	.35	.52	.69	.59	

Note. The diagonal entries are fidelity coefficients. Factors are presented in the same order as the scales. JDS = Job Diagnostic Survey.

Instrument. The JDS includes 15 items that measure the five core job characteristics of task identity, task significance, skill variety, autonomy, and feedback. Each of these core dimensions was measured by three 7-point Likert-type items. The internal consistency reliabilities of the scales, using the entire sample of *N* = 6,930, ranged from .71 (skill variety and feedback from the job itself) to .59 (task identity).

Results

Eigenvalues of the reduced correlation matrix for one through seven factors were 3.68, 0.87, 0.64, 0.42, 0.38, 0.20, and 0.02, respectively. The corresponding parallel analysis eigenvalues were 0.17, 0.13, 0.10, 0.08, 0.06, 0.05, and 0.03. These eigenvalues indicate that there are six factors underlying the JDS items. The maximum likelihood goodness-of-fit measure also clearly indicated that more than five factors were required to model the correlations among the JDS items.

Table 1 presents the end result of the maximum likelihood factor analyses. The chi-square goodness-of-fit measure for this solution is 162.58 with 58 *dfs*. The six-factor solution for the principal axes factor analysis using the DAPPFR rotation closely resembles this solution and therefore is not presented. In both of the solutions, five of the factors correspond to the JDS a priori pattern. Skill variety items load on Factor 1, task identity items load on Factor 2, task significance items load on Factor 4, autonomy items load on Factor 5, and feedback items load on Factor 6. In addition, all items requiring reverse scoring load on a single factor (Factor 3). The same reversed scoring factor was obtained when we extracted five factors.

The factor correlation matrix is also included in Table 1. As anticipated, the JDS factors are moderately intercorrelated: The average factor correlations between JDS factors is .42. Conceptually, these relations may result from the existence of a common latent construct that is directly linked to manifest characteristics of the task (Hulin & Roznowski, 1985). How-

ever, common method variance cannot be overlooked as a possible cause of these correlations.

The average correlation between the artifact factor (or *arti-factor*) and the JDS factors is .12 after the DAPPFR rotation and .17 in the maximum likelihood factor analysis solution. These coefficients are expected to be small because the arti-factor does not measure a latent construct related to the technological aspects of jobs. In addition, the small arti-factor–JDS factor coefficients suggest that the average correlation of .42 among the JDS factors may not be due entirely to method variance. However, as one reviewer of this article pointed out, the factor inter-correlations may also be caused, in part, by other artificial sources.

Job Diagnostic Survey scale–factor correlations were computed using the method presented by Drasgow and Miller (1982). The fidelity coefficients obtained from the restricted maximum likelihood factor analyses solution are presented in the main diagonal of Table 2. The range of the coefficients extends from .75 to .85 for the JDS scales; it is .71 for the arti-factor. The fidelity coefficients for principal axes factor analysis and DAPPFR rotation were very similar.

According to Drasgow and Miller (1982), scale–factor coefficients above .90 may be necessary for construct validation research. This criterion is very difficult to attain, however, for scales as short as the JDS scales. For some types of research, coefficients in the .80s may be acceptable as long as the variance in scales not accounted for by the factors they measure is due to random measurement error. In Study 1, some of the fidelity coefficients fell below the .80 criterion and the scales contained systematic errors caused by the measurement artifact. It is evident that some revision of items requiring reverse scoring is necessary.

Study 2

Purpose

The second study was performed to replicate Study 1 with an independent sample.

Method

Data for Study 2 consisted of subsamples from Oldham et al.’s (1978) JDS data base. This sample (*N* = 565) consisted of 132 service workers,

Table 3
Revised Job Diagnostic Survey Items

Scale	Item
Autonomy	The job gives me a chance to use my personal initiative and judgment in carrying out the work.
Task Identity	The job is arranged so that I can do an entire piece of work from beginning to end.
Skill Variety	The job requires me to use a number of complex or high-level skills.
Task Significance	The job itself is very significant and important in the broader scheme of things.
Feedback	After I finish a job, I know whether I performed well.

Table 4
DAPPPFR Rotation of the Six-Factor Principal Axes Factor
Analysis Solution for the Revised JDS

Item	Factor					
	1	2	3	4	5	6
Autonomy						
1					.52	
2			.35		.43	.09 ^a
3					.66	
Task Identity						
4	.60					-.21
5	.77					.07 ^a
6	.74					
Skill Variety						
7			.74			
8			.49			.22
9			.76			.09 ^a
Task Significance						
10		.72				
11		.68				
12		.64		.23		.05 ^a
Feedback						
13				.60		
14				.64		
15				.57	.33	.03 ^a

Factor correlation matrix						
1	—					
2	-.11	—				
3	.30	.28	—			
4	.22	.27	.32	—		
5	.42	.15	.47	.33	—	
6	.10	.09	-.18	.25	-.15	—

Note. All loadings less than .20 are omitted except for the loadings of the rewritten items on Factor 6. DAPPPFR = direct artificial personal probabilities factor rotation; JDS = Job Diagnostic Survey.
^a This item is rewritten so that reverse scoring is not necessary.

161 bench workers, and 272 workers classified as *other*. These job categories were selected so that they were mutually exclusive with the categories used in Study 1. Of the sample 71% were male; 68% of the workers were between the ages of 20 and 40 years; 50% had a high school degree or less education; and 35% had some technical or college education but did not receive a degree.

Results

The first seven eigenvalues of the reduced correlation matrix for the second sample are 3.32, 0.73, 0.67, 0.64, 0.39, 0.23, and 0.13, respectively. The corresponding parallel analysis eigenvalues are 0.32, 0.25, 0.20, 0.16, 0.13, 0.09, and 0.06. Once again, six factors seem to be plausible based on the parallel analysis results.

Principal axes factor analysis with a DAPPPFR rotation and restricted maximum likelihood factor analysis were computed for six dimensions. The results of the two analyses were very similar to our findings in Study 1 and consequently are not presented. A sixth factor again emerged that was defined by the five reverse-scored items.

Study 3

Purpose

In Studies 1 and 2 it was determined that an arti-factor is needed to explain statistically the intercorrelations of JDS items. Moreover, the arti-factor is clearly detrimental to the measurement accuracy of the JDS scales as indicators of the a priori factors because it systematically affects scale scores.

In Study 3 we attempted to improve the measurement properties of the JDS scales by revising the items requiring reverse scoring. We attempted to rewrite these items in ways that maintained their original meanings yet did not require reversed scoring. The revised JDS was administered to a third sample of workers and the responses were factor analyzed.

Method

Subjects. Data for Study 3 were collected from 94 female and 40 male employees of a printing plant in central Illinois. All of the employees working in the customer information/service, documentation, mailing service/records, distribution, order processing, printing, and bindery departments were asked to participate. Of the employees who responded, 87% were white, 47% were married, 56% had up to a high school education, and 31% had 1 to 3 years of college education. The average age of the respondents was 31, the average tenure with the organization was between 1 and 3 years, and the respondents worked an average of 38 hr per week.

Instrument. A revised version of the JDS was developed after consultation with the instrument's original authors. The format of the revised survey was the same as the original JDS. Only items requiring reverse scoring in their original form were rewritten. The five rewritten items are presented in Table 3. Note that this listing does not reflect the format of the JDS.

Administrative procedures. The written questionnaires were distributed to employees by the supervisor of each department. Employees were informed that their participation was voluntary, their replies would be kept anonymous, and employees of the company would only see results based on aggregated data. Employees were given 1 week to complete the questionnaire. The return rate across all departments was approximately 65%.

Results

The first seven eigenvalues of the reduced correlation matrix are 3.72, 1.17, 0.83, 0.72, 0.52, 0.15, and 0.07, respectively. The corresponding parallel analysis eigenvalues are 0.76, 0.59, 0.48, 0.39, 0.31, 0.24, and 0.14. These eigenvalues indicate that five factors underlie the revised JDS. We, nonetheless, examined the six-factor solution in order to verify that our revisions eliminated the arti-factor.

The six-factor solution that resulted from the DAPPPFR rotation is presented in Table 4. It clearly indicates that we have extracted too many factors. Notice that all loadings of the rewritten items on Factor 6 are nearly zero.

Given the lack of an arti-factor in the principal axes factor solution, the five a priori core job dimensions should now appear as distinct dimensions in the five-factor solution. Principal axes factor analysis and both unrestricted and restricted maximum likelihood factor analysis were used to obtain five-factor structures. The goodness-of-fit measures for the unrestricted

Table 5
Five-Factor Maximum Likelihood Factor Structure for the Revised Job Diagnostic Survey

Item	Factor				
	1	2	3	4	5
Autonomy					
1					42 (10)
2			29 (09)	29 (09)	37 (10)
3					89 (11)
Task Identity					
4	62 (09)				
5	79 (08)				
6	80 (08)				
Skill Variety					
7			81 (10)		
8			32 (10)	31 (10)	
9			66 (10)	29 (10)	
Task Significance					
10		70 (09)			
11		71 (09)			
12		68 (09)		25 (08)	
Feedback					
13				61 (09)	
14				85 (08)	
15				56 (09)	

Factor correlation matrix					
1	—				
2	−13 (11)	—			
3	25 (11)	16 (11)	—		
4	26 (10)	26 (11)	24 (12)	—	
5	42 (10)	13 (11)	37 (11)	37 (10)	—

Note. All factor loadings fixed at zero are omitted; decimal points are omitted; standard errors of parameter estimates are in parentheses.

and restricted solutions, respectively, are $\chi^2(40) = 44.04$, and $\chi^2(74) = 86.90$. Both measures indicate a good fit.

The results for the restricted maximum likelihood factor analysis are presented in Table 5. The DAPFPR solution was very similar. As expected, the hypothesized JDS structure is clearly evident in Table 5. Factors 1 through 5 can be identified as the job dimensions of Task Identity, Task Significance, Skill Variety, Feedback, and Autonomy, respectively. The factor intercorrelations ranged from $-.13$ to $.42$ in the maximum likelihood factor analysis solution.

To check the measurement accuracy of the revised scales, fidelity coefficients were computed. Table 2 presents the Study 3 fidelity coefficients that resulted from the restricted maximum likelihood factor analysis. The fidelity coefficients for the Task Identity, Task Significance, Skill Variety, Autonomy, and Feedback factors are all above $.80$. In sum, it appears that the fidelity coefficients for the revised JDS scales are reasonably high and we can conclude that the JDS scales are measuring their underlying constructs with accuracies that are adequate for theoretical research.

Discussion

Since the introduction of the Job Diagnostic Survey, numerous studies have attempted, with limited success, to obtain empirical support for the hypothesized five-factor structure. Reasons for the lack of success in this area of research were not

apparent until the Harvey et al. (1985) study. They suggested that reverse-scored items were a major source of the inconsistencies. In the present study we provide confirmation that the reversed scored items are indeed the source of the problem. It is interesting that this arti-factor can also be identified in the factor structures presented by Dunham (1976), Dunham et al. (1977), and to a lesser extent, Pokorney et al. (1980) and Fried and Ferris (1986). Given the diversity of the samples used in these factor analyses, it seems safe to conclude that the reverse-scored items have caused the difficulties in factoring the JDS. Ironically, Hackman and Oldham (1974, 1975) deliberately incorporated reverse-scored items into the survey to minimize response bias. Unfortunately, their effort backfired and seems to have caused a substantial amount of mischief.

Identification of this measurement artifact may have been delayed, until recently, for several reasons. First, few studies have extracted more than five factors for the original JDS items. In many cases, such as in the Dunham et al. (1977) four-factor solution, the arti-factor seems to be stronger than some of the JDS factors. When the arti-factor appeared in a solution, it may have misled researchers to believe that they had overfactored and, consequently, they may not have examined solutions in higher dimensionalities.

In addition, identification may have been delayed because the salience or the “strength” of the arti-factor may vary as a result of differences in reading comprehension (Green et al., 1979),

education and position level (Fried & Ferris, 1986), and attentiveness to negatively worded items (Schmitt & Stults, 1985).

Fried and Ferris's (1986) results suggest that the arti-factor is "weaker" for workers with more education or higher reading comprehension ability. They obtained good approximations to the a priori structure only in subsamples with more education and higher position levels. Such workers may be more adept at reversing the negatively scored items and may therefore respond using the same criteria that are used on other items. Workers with lower reading abilities and less education may have more difficulty in reversing items mentally and therefore may respond incorrectly to the items, thus creating the arti-factor. It is interesting to note that Schmitt and Stults (1985) found that "negative" factors appear when as few as 10% of the subjects fail to notice that some items are reverse scored.

Our approach to revising the JDS was to replace reverse-scored items with new items that did not have to be reverse scored. Some of the results presented by Harvey et al. (1985) suggest that it may also be necessary to use a single response format to eliminate all measurement artifact factors. However, we found no evidence of an arti-factor due to response format in Study 3. Of course, additional research with large samples collected from diverse organizations is needed before we can conclude unequivocally that the measurement artifacts have been eliminated. At present it seems appropriate to conclude that the new JDS items have substantially improved measurement properties and the new scales should be used in future research concerned with task characteristics.

Although the measurement artifact seems to have been eliminated, investigations are still needed to better understand why reverse phrasing of some items can lead to artifact factors. It is possible that the arti-factor is simply a consequence of the lack of salience of the five items written in a different, reverse-scored format. A reviewer suggested that it may be possible to eliminate the arti-factor by explicitly pairing items that are worded in opposite directions. To study this hypothesis for the JDS, items could be written and paired with current JDS items. The relative salience argument would be supported if the arti-factor did not appear in a factor analysis of the augmented scale.

In any case, a revised JDS now exists that appears to adequately measure the five task dimensions explicated by Hackman and Oldham (1974, 1975). We are now able to go beyond considerations of the psychometric adequacy of the JDS and study the conceptual role of task characteristics in organizational behavior.

References

- Drasgow, F., & Miller, H. E. (1982). Psychometric and substantive issues in scale construction and validation. *Journal of Applied Psychology*, 67, 268-279.
- Dunham, R. B. (1976). The measurement and dimensionality of job characteristics. *Journal of Applied Psychology*, 61, 404-409.
- Dunham, R. B., Aldag, R. J., & Brief, A. P. (1977). Dimensionality of task design as measured by the Job Diagnostic Survey. *Academy of Management Journal*, 20, 209-223.
- Fried, Y., & Ferris, G. R. (1986). *The dimensionality of job characteristics: Some neglected issues*. Manuscript submitted for publication.
- Green, S. B., Armenakis, A. A., Marbert, L. D., & Bedeian, A. G. (1979). An evaluation of the response format and scale structure of the Job Diagnostic Survey. *Human Relations*, 32, 181-188.
- Hackman, J. R., & Oldham, G. R. (1974). *The Job Diagnostic Survey: An instrument for the diagnosis of jobs and the evaluation of job redesign projects* (Report No. 4). New Haven, CT: Yale University, Department of Administration Science.
- Hackman, J. R., & Oldham, G. R. (1975). Development of the Job Diagnostic Survey. *Journal of Applied Psychology*, 60, 159-170.
- Harvey, R., Billings, R., & Nilan, K. (1985). Confirmatory factor analysis of the Job Diagnostic Survey: Good news and bad news. *Journal of Applied Psychology*, 70, 461-468.
- Hulin, C. L., & Roznowski, M. (1985). Organizational technologies: Effects on originators' characteristics and individual responses. In L. L. Cummings & B. M. Staw (Eds.), *Research in organizational behavior* (Vol. 7, pp. 39-85). Greenwich, CT: JAI Press.
- Humphreys, L. G., & Montanelli, R. G., Jr. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research*, 10, 193-205.
- Kroonenberg, P. M., & Lewis, C. (1982). Methodological issues in the search for a factor model: Exploration through confirmation. *Journal of Educational Statistics*, 7, 69-89.
- Mowday, R. T., Porter, L. W., & Steers, R. M. (1982). *Employee-organization linkages: The psychology of commitment, absenteeism, and turnover*. New York: Academic Press.
- Oldham, G. R., Hackman, J. R., & Stepina, L. P. (1978). *Norms for the Job Diagnostic Survey* (Tech. Rep. No. 16). New Haven, CT: Yale University, School of Organization and Management.
- Pierce, J. L., & Dunham, R. B. (1978). The measurement of perceived job characteristics: The Job Diagnostic Survey versus the Job Characteristic Inventory. *Academy of Management Journal*, 21, 123-128.
- Pokorney, J. J., Gilmore, D. C., & Beehr, T. A. (1980). Job Diagnostic Survey dimensions: Moderating effect of growth needs and correspondence with dimensions of the Job Rating Form. *Organizational Behavior and Human Performance*, 26, 222-237.
- Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, 9, 367-373.
- Tucker, L. R., & Finkbeiner, C. T. (1981). *Transformation of factors by artificial personal probability functions* (Research Rep. No. RR-81-58). Princeton, NJ: Educational Testing Service.
- Turner, A. N., & Lawrence, P. R. (1965). *Industrial jobs and the worker*. Boston, MA: Harvard Graduate School of Business Administration.

Received February 28, 1986

Revision received May 19, 1986 ■

Employee Reactions to Workspace Characteristics

Greg R. Oldham

University of Illinois at Urbana-Champaign

Yitzhak Fried

Wayne State University

We investigated the independent and joint effects of four workspace characteristics (social density, room darkness, number of enclosures, and interpersonal distance) on three employee reactions: turnover, satisfaction, and withdrawal from the office during discretionary periods. A total of 109 clerical employees from 19 offices of a large university participated in the research. Results showed that the independent and joint effects of the workspace characteristics accounted for 24% of the variance in employee turnover, 31% of the variance in work satisfaction, and 34% of the variance in discretionary withdrawal. Moreover, the four-way interaction term involving the workspace characteristics contributed significantly to each of the reaction measures, suggesting that employees were most likely to withdraw from offices and to experience dissatisfaction when the following conditions were present: the office was rated as dark, few enclosures surrounded employees' work areas, employees were seated close to one another, and many employees occupied the office. The implications of the findings for future research on workspace design are discussed.

Numerous studies in social and environmental psychology have demonstrated that characteristics of the physical environment (e.g., density and partitions) have a substantial effect on the behavior and attitudes of individuals (cf. Desor, 1972; Paulus, Annis, Seta, Schkade, & Matthews, 1976; Walden & Forsyth, 1981). Unfortunately, very little is known about the impact that characteristics of the work setting or workspace might have on the reactions of employees (Becker, 1981; Sundstrom, 1986; Wineman, 1982). The purpose of the present investigation was to address this problem. Specifically, we examined the independent and joint effects of several workspace characteristics on a variety of employee reactions (e.g., turnover and satisfaction).

A number of theories in social and environmental psychology have been developed to explain the effects of characteristics of the physical environment on individuals' reactions (e.g., interference theory, Schopler & Stockdale, 1977; intensification theory, Freedman, 1975; control-loss theory, Baron & Rodin, 1978). Prominent among these is overstimulation theory (Desor, 1972; Saegert, 1978). In this theory it is argued that certain features of the physical environment contribute to excessive stimulation, which leads to a psychological state of stimulus overload. Overstimulation can derive from too many people, too many interactions, too close a proximity of others, and small amounts of space (Paulus, 1980). Individuals are expected to react to excessive stimulation both behaviorally and attitudinally (Paulus, 1980; Vine, 1981). In the context of a work organization, employees might physically withdraw from

an overstimulating work environment and experience dissatisfaction with the work they do in that environment. The latter is a possibility because individuals may have difficulty focusing and concentrating on their work when they are overly stimulated by external conditions (Oldham & Rotchford, 1983).

Four workspace characteristics derived from previous research in the area of social and environmental psychology (e.g., Baum & Davis, 1976; Desor, 1972; Paulus et al., 1976) are investigated in this study: social density, room darkness, number of enclosures, and interpersonal distance. Each of these characteristics might contribute to stimulus overload and therefore influence employee reactions. The characteristics are described next in more detail.

Social density refers to the total number of individuals in a particular area (e.g., an office or a store)—regardless of the amount of space available in that area (Hayduk, 1983; Paulus, 1980). In general, previous research on this variable suggests that individuals often respond negatively to socially dense conditions (see Paulus, 1980, and Sundstrom, 1978, for reviews). Specifically, research has shown that individuals in high-social-density conditions feel more crowded and are less attracted to others than individuals in conditions of low density (Bharucha-Reid & Kiyak, 1982; Marshall & Heslin, 1975). In addition, other studies have shown that social density is positively correlated with turnover intentions and negatively related to individuals' job satisfaction and task performance (Dean, Pugh, & Gunderson, 1975; Paulus et al., 1976; Sundstrom, Burt, & Kamp, 1980).

Room darkness refers to the overall darkness of a setting (e.g., an office). Illumination levels and wall colors might contribute substantially to this dimension.

Previous research has demonstrated that individuals perceive dark settings as smaller and as more crowded than light settings (Baum & Davis, 1976; Mandel, Baron, & Fisher, 1980; Schiffrinbauer, Brown, Perry, Shulack, & Zanzola, 1977). For example, Baum and Davis (1976) found that individuals rated rooms painted dark colors, as smaller and more crowded than

The authors express great appreciation to Marcia Kassner, Nancy Rotchford, and Christina Shalley for their assistance in data collection and analysis, and to Carol Kulik, Keith Murnighan, and Joseph Porac for their helpful comments on earlier drafts of this article.

Correspondence concerning this article should be addressed to Greg R. Oldham, Department of Business Administration, 350 Commerce West, University of Illinois, 1206 South Sixth, Champaign, Illinois 61820.

rooms painted light colors. Other studies have shown that individuals experience less workspace, work and social satisfaction, and perform at lower rates in dark, as opposed to light, rooms (Barnaby, 1980; Brill, 1984; Oldham & Rotchford, 1983; Smith, 1978).

The connection between overstimulation theory and room darkness requires some elaboration. Individuals may experience considerable spatial restriction in dark as opposed to light rooms because the former are typically perceived as smaller and less spacious than the latter (Baum & Davis, 1976). This perceived restriction of space may contribute substantially to the experience of overstimulation (Paulus, 1980).

Number of enclosures refers to the number of walls or partitions surrounding an individual's work area (e.g., desk). It is probable that physical barriers surrounding a work area limit to some degree, the unwanted or unexpected intrusions that contribute to overstimulation (Archea, 1977).

Previous research has demonstrated that number of enclosures is positively correlated with individuals' job performance, workspace satisfaction, and experienced privacy, and negatively associated with individuals' perceptions of crowding (Desor, 1972; Sundstrom et al., 1980; Sundstrom, Town, Brown, Forman, & McGee, 1982). In addition, Oldham and Rotchford (1983) showed that employees were most likely to withdraw from an office during discretionary periods when there were few partitions surrounding their individual work areas.

Interpersonal distance refers to the distance between an individual and the nearest person in a given area (Paulus et al., 1976; Sundstrom et al., 1980). Previous research has shown that individuals feel more crowded, confined, distracted, and uncomfortable when there is little distance between them and another person than when there is a substantial distance (Walden & Forsyth, 1981; Worchel & Teddlie, 1976). Furthermore, other studies have demonstrated that individuals perform at lower levels in close as opposed to far distance conditions (Paulus et al., 1976; Sundstrom et al., 1980).

In summary, previous research suggests that individuals react negatively to the following characteristics of the physical environment: high social density, dark rooms, close interpersonal distance, and few partitions or enclosures. In the present investigation we extended this earlier work and examined associations between measures of these four characteristics and three employee reactions: turnover, withdrawal from the work setting during discretionary periods, and work satisfaction. In addition, we examined the possibility that the four workspace characteristics interact with one another to influence these employee reactions, as suggested by several studies in social and environmental psychology (e.g., Bharucha-Reid & Kiyak, 1979, 1982; Lange, Mueller, & Donnerstein, 1979; Paulus et al., 1976; Worchel & Teddlie, 1976).

Based on the research reviewed here, we tested the following basic hypotheses:

Hypothesis 1. Each of the workspace characteristics relates positively to turnover and discretionary withdrawal and negatively to work satisfaction. For example, it is predicted that the higher the social density, the higher the turnover and withdrawal and the lower the satisfaction.

Hypothesis 2. Employees experience the most stimulus overload and therefore react most negatively when all four

workspace characteristics are present simultaneously (i.e., when the work setting is dark, when social density is high, when there are few enclosures, and when there is little distance between employees).

Method

Research Setting and Subjects

The research was conducted in 19 offices of a large midwestern university. Data were collected from 114 full-time employees; however, missing data reduced the final sample to 109. A total of 93% ($n = 101$) of the participants were women. The number of participants from each of the offices ranged from 2 to 20. All of the employees worked in clerical jobs, which were at approximately the same level in the university and involved very similar duties. The mean education level was "some college experience." The mean tenure level was 39 months.

Procedure

Data were collected on site by the first author or an associate. A questionnaire was administered to groups of employees (ranging from 2 to 15 at a time). The questionnaire included items that measured withdrawal during discretionary periods and work satisfaction. Before completing the questionnaire, employees were told about the nature and purpose of the research and were given the option of not participating. It was also emphasized that all information obtained would be held in confidence, and that no one in the office would have access to individual responses. Employees were told that it was desirable to have names on questionnaires both for research purposes and so that a feedback report summarizing participants' responses could be provided to them. All of the employees agreed to complete the questionnaire and to provide their names.

The researchers evaluated the four workspace characteristics in each of the 19 offices typically during the noon hour, when employees were absent. Turnover data were obtained from office records.

Measures

Workspace characteristics. Each of the characteristics was measured as follows.

1. *Social density:* This measure was created by calculating the total number of employees who worked in each of the 19 offices. The more employees in the office, the higher the social density (Hayduk, 1983; Paulus, 1980).

2. *Darkness:* The researchers rated (a) the color of the office walls and (b) the level of office illumination on 3-point scales ranging from light to dark. Scores on the two measures were then averaged to form an office darkness index ($\alpha = .59$). The higher the score on this measure, the darker the office.

3. *Number of enclosures:* This measure was created by counting the number of walls and partitions that surrounded each employee's desk. Doors that could be closed were counted as walls. Scores on the enclosures measure ranged from 0 to 4. This value was then reverse scored; thus, the higher score, the fewer the enclosures.

4. *Interpersonal distance:* This was calculated by measuring the distance between the employee's desk and the nearest co-worker's desk. Specifically, the number of inches between the employee's desk center and the co-worker's desk center was calculated (Sundstrom et al., 1980). This value was then reverse scored; thus, the higher the score, the closer the nearest co-worker.

Two of the workspace characteristics (i.e., social density and darkness) are by definition characteristics of the overall office setting. Therefore, there were only 19 separate scores for each of these two workspace

Table 1
Means, Standard Deviations, and Intercorrelations Among All Variables

Variable	M	SD	1	2	3	4	5	6	7
1. Density	12.7	9.4	—						
2. Darkness	1.7	0.5	.24*	—					
3. Enclosures	2.0	0.9	.00	.27*	—				
4. Distance	117.6	65.3	-.03	.19*	.46*	—			
5. Office turnover	0.3	0.5	.06	.26*	.04	-.03	—		
6. Discretionary withdrawal	0.5	0.5	.12	.29*	.27*	.17	.04	—	
7. Work satisfaction	3.1	1.3	.00	.23*	.05	.07	.08	.11	—

Note. $N = 109$. * $p < .05$, two-tailed.

variables—one density score and one darkness score for each of the offices included in the research. Consistent with methods used in previous research (e.g., Oldham & Hackman, 1981; Oldham & Rotchford, 1983; Pierce, 1979; Pierce, Dunham, & Cummings, 1984; Rousseau, 1978), all participants from a given office were assigned these “macro”-office scores. Roberts, Hulin, and Rousseau (1978) suggested that such an approach is appropriate when individuals are homogeneous with respect to a particular macrovariable. With regard to the variables examined in this study, we argue that all employees in the same office were exposed to the darkness of that office, as well as to the social density of that office. Therefore, the employees in the offices were considered homogeneous with respect to these macromasures of the physical environment.

Reactions. Each of the reactions was measured as follows.

1. Office turnover: Approximately 24 months after the workspace data were collected, turnover data were obtained from office records. Employees who remained with the office were assigned a score of 0; those who left were assigned a score of 1. The turnover rate was approximately 34%. Interviews with office management indicated that the turnover measured in this research was voluntary in nature. And results of a series of t tests indicated that “leavers” did not differ significantly ($p < .05$, two-tailed) from “stayers” on the following personal characteristics: sex ($t = 0.36$), education ($t = 0.91$), and tenure ($t = 1.39$).

2. Discretionary withdrawal: Employees indicated on the questionnaire where they typically spent their coffee break periods. These responses were then analyzed by the researchers. Employees who remained in their offices during breaks were assigned a score of 0; those who went elsewhere were assigned a score of 1.

3. Work satisfaction: Nine items from the “general” and “growth” satisfaction sections of the Job Diagnostic Survey (Hackman & Oldham, 1975) were averaged to create a measure of work satisfaction ($\alpha = .88$). Responses to this measure were reverse scored: the higher the score, the lower the satisfaction.

Statistical Analyses

Two types of analyses were used to test the hypotheses. To test the first hypothesis, zero-order correlations between the workspace characteristics and the reaction measures were examined. For the second hypothesis, hierarchical regression analyses were used to examine the amount of unique criterion variance explained by the interactions among the workspace characteristics. In these analyses, sets of variables were introduced into the regression equations in a stepwise fashion. Specifically, the four workspace characteristics were introduced into the equations, followed by the 6 two-way interactions, 4 three-way interactions, and the 1 four-way interaction. We emphasize interpretation of the increased squared multiple correlation (R^2) that results from including a particular predictor in the regression equation as an indication of its importance, because of problems of multicollinearity and instability of regres-

sion coefficients (Hom, Griffeth, & Sellaro, 1984; Miller, Katerberg, & Hulin, 1979).

Results

Relations Among the Measures

Correlations among the workspace and reaction measures are reported in Table 1. The table shows that there were no significant correlations among the three reaction measures. Correlations among the workspace characteristics were generally positive and statistically significant. However, only one of the relations was of substantial magnitude—that between interpersonal distance and number of enclosures.

Correlations between the workspace characteristics and the reaction measures are also shown in Table 1. Results provide only partial support for Hypothesis 1. The number-of-enclosures measure related significantly to discretionary withdrawal—the fewer barriers around an employee’s desk, the more likely he or she was to take coffee breaks outside the boundaries of the office. Also, office darkness correlated significantly with the turnover, discretionary withdrawal, and satisfaction measures. Employees from dark offices were more likely to withdraw and to express dissatisfaction with their work than were employees from light offices. Results involving the social density and distance characteristics did not support the hypothesis. There were no significant correlations between these characteristics and the reaction measures.

Hierarchical Regression Analyses

To test the contributions of the interactions among the workspace characteristics to the explanation of the reaction measures, a hierarchical regression analysis was performed on each of the three dependent variables. Results are summarized in Table 2. The amount of variance accounted for in the dependent variables by the main effects alone ranged from 8% for turnover to 14% for discretionary withdrawal. However, the interaction terms made statistically significant contributions to the explanation of each of the three reaction measures. Specifically, the Density \times Enclosures and Darkness \times Distance \times Enclosures interactions contributed significantly to the explanation of turnover; the Darkness \times Enclosures interaction contributed to the explanation of satisfaction; and the Density \times Darkness, Darkness \times Enclosures \times Distance, and Density \times

Table 2
Summary of Hierarchical Regression Analyses

Measure	R ²	ΔR ²	Measure	R ²	ΔR ²	Measure	R ²	ΔR ²
Office turnover			Discretionary withdrawal			Work satisfaction		
A	.07*	.07*	A	.09*	.09*	A	.07*	.07*
B	.08*	.01	C	.13*	.04*	B	.08*	.01
C	.08*	.00	D	.14*	.01	D	.09*	.01
D	.08	.00	B	.14*	.00	C	.09*	.00
C × D	.12*	.04*	A × D	.18*	.04*	A × C	.18*	.09*
A × B	.14*	.02	B × D	.19*	.01	A × B	.20*	.02
A × D	.14*	.00	A × C	.20*	.01	A × D	.21*	.01
B × C	.15*	.01	C × D	.20*	.00	B × D	.21*	.00
A × C	.15	.00	B × C	.20*	.00	B × C	.22*	.01
B × D	.15	.00	A × B	.20*	.00	C × D	.22*	.00
A × B × C	.19*	.04*	A × B × D	.27*	.07*	B × C × D	.23*	.01
A × B × D	.20*	.01	A × B × C	.31*	.04*	A × B × C	.24*	.01
A × C × D	.20	.00	A × C × D	.31*	.00	A × B × D	.25*	.01
B × C × D	.20	.00	B × C × D	.31*	.00	A × C × D	.25*	.00
A × B × C × D	.24*	.04*	A × B × C × D	.34*	.03*	A × B × C × D	.31*	.06*

Note. *n* = 109. A = darkness; B = distance; C = enclosures; D = density.
* *p* < .05.

Darkness × Distance interactions contributed significantly to the explanation of the discretionary withdrawal measure. Finally, the four-way interaction term made a statistically significant unique contribution to each of the three reaction measures. This interaction term explained 4% of the variance in turnover, 3% in discretionary withdrawal, and 6% in work satisfaction.

To interpret these interactions, the beta weights for the interaction terms were examined. In all cases, the signs of the beta weights were positive, indicating that employees reacted negatively when the characteristics involved in the interaction were present simultaneously at high levels. Thus, for the significant Enclosures × Density interaction, turnover was higher when there were few enclosures in an office occupied by many employees. For the significant Darkness × Enclosures interaction, satisfaction was lower when the office was dark and few enclosures were present. And the results for the four-way interaction terms indicate that turnover and discretionary withdrawal were highest and work satisfaction lowest when all four of the physical characteristics were present simultaneously at high levels (i.e., few enclosures, close interpersonal distance, etc.).¹ These latter results provide general support for Hypothesis 2.

Discussion

The results of this study suggest that the physical characteristics of a work environment can have an impact on the behavioral and attitudinal reactions of employees. Specifically, the independent and joint effects of four workspace characteristics (i.e., social density, darkness, enclosures, and interpersonal distance) accounted for 24% of the variance in employee turnover, 34% of the variance in withdrawal during discretionary periods, and 31% of the variance in work satisfaction. Moreover, the four-way interaction term involving the workspace characteristics contributed significantly to each of the reaction measures. Employees were most likely to withdraw from offices and to

experience dissatisfaction with their work when the following conditions were present simultaneously: the office was rated as dark, few enclosures surrounded employees' work areas, employees were seated close to one another, and many employees occupied the office. If any one of these conditions was missing or absent, employees reacted less negatively.

Although this study contributes to our understanding of the impact of characteristics of the physical environment on employee reactions, a number of issues deserve research attention in the future. One of these involves the approach used to measure the workspace characteristics. In the current study, only one measure of each workspace characteristic was obtained, and it is not clear that these measures converge with possible alternative measures (e.g., ratings of density made by an observer, or darkness scores obtained from a light measuring apparatus). Future research should use multiple measures to assess each of the workspace characteristics and should examine the associations among these measures. Of particular interest are relations between office employees' ratings of the workspace characteristics and ratings obtained from some other source (e.g., a mechanical device or an observer). Although we know of no previous study that examined such relations, previous research in the area of work design suggests that there is likely to be considerable agreement between such ratings (see Fried, 1985, for a review). For example, several studies have shown that ratings of job characteristics by incumbents converge substantially with ratings made by external observers (Brass, 1981;

¹ To further investigate these four-way interactions, several subgroup analyses were conducted. In these analyses, three of the workspace characteristics were divided at the median and then combined to form eight separate subgroups (e.g., high density, high darkness, low enclosures). Correlations between the fourth workspace characteristic and the three reaction measures were next calculated for each of the subgroups. Results were consistent with those obtained in the regression analyses. Details of these results are available from the authors on request.

Hackman & Lawler, 1971; Hackman & Oldham, 1975). Moreover, research has also shown that observers' ratings of job characteristics explain about as much variance in incumbents' reactions as do ratings made by the incumbents themselves (Jenkins, Glick, & Gupta, 1983; Oldham, Hackman, & Pearce, 1976; Stone & Porter, 1978). Despite this evidence from the area of work design, research on the physical environment that includes both employee ratings of workspace characteristics and ratings from other sources could be very informative and should be conducted in the near future.

In addition, future studies might examine the independent and joint effects of the workspace characteristics on other employee reactions (e.g., performance, absenteeism, and tardiness). Research is also needed on the effects of workspace characteristics not investigated in this study (e.g., distance to supervisors, distance to walkways, window placement, and the number of rooms available for private discussion within the office setting). Finally, the specific components of several of the workspace characteristics examined in this research are deserving of research attention in the future. For example, it would be worthwhile to determine the impact of the nature of the enclosures (e.g., clear plastic partitions or solid, soundproof walls) and of the sources of office lighting (e.g., wall colors, natural lighting, artificial lighting).

Research is also needed on the impact of individual differences on the relations between workspace characteristics and employee responses. In the current study, we argued that certain workspace characteristics can contribute to excessive stimulation that, in turn, can lead to a number of adverse reactions. It may be that some individuals are better able to cope with and manage this excessive stimulation than are others. In this regard, Mehrabian (1977) has developed a stimulus screening measure that distinguishes between people who are able to cope with numerous inputs and stimuli (screeners) and those who are not (nonscreeners). It may be that screeners react less negatively to certain workspace characteristics (e.g., darkness and social density) than nonscreeners. Some support for this possibility is found in a recent study of the impact of stimulus screening on students' responses to dense dormitory environments (Baum, Calesnick, Davis, & Gatchel, 1982).

Finally, one should use caution when attempting to generalize from the results of this investigation. The study focused only on clerical employees, 93% of whom were women. It may be that people in different jobs (e.g., managerial or research and development) require different workspace characteristics to successfully complete their work and therefore respond differently to characteristics of the physical environment than do individuals holding clerical jobs. And it may be that women react differently than men to certain workspace characteristics. In support of the latter possibility, a number of studies suggest that men react more negatively than women to high-spatial-density conditions (i.e., conditions in which limited space is available) (Baum & Koman, 1976; Epstein & Karlin, 1975; Paulus et al., 1976). Research is now needed to determine if men and women employees respond differently to several characteristics of the workspace.

References

- Archea, J. (1977). The place of architectural factors in behavioral theories of privacy. *Journal of Social Issues*, 33, 116-137.
- Barnaby, J. F. (1980). Lighting for productivity gains. *Lighting Design and Application*, 2, 20-28.
- Baron, R. M., & Rodin, J. (1978). Personal control as a mediator of crowding. In A. Baum, J. Singer, & S. Valins (Eds.), *Advances in environmental psychology: Vol. 1. The built environment* (pp. 145-190). Hillsdale, NJ: Erlbaum.
- Baum, A., Calesnick, L. E., Davis, G. E., & Gatchel, R. J. (1982). Individual differences in coping with crowding: Stimulus screening and social overload. *Journal of Personality and Social Psychology*, 43, 821-830.
- Baum, A., & Davis, G. E. (1976). Spatial and social aspects of crowding perception. *Environment and Behavior*, 8, 527-545.
- Baum, A., & Koman, S. (1976). Differential response to anticipated crowding: Psychological effects of social and spatial density. *Journal of Personality and Social Psychology*, 34, 526-536.
- Becker, F. D. (1981). *Workspace: Creating environments in organizations*. New York: Praeger.
- Bharucha-Reid, R., & Kiyak, H. A. (1979). The concept of dissonance and too much personal space. *Journal of Nonverbal Behavior*, 4, 123-125.
- Bharucha-Reid, R., & Kiyak, H. A. (1982). Environmental effects on affect: Density, noise and personality. *Population and Environment*, 5, 60-72.
- Brass, D. J. (1981). Structural relationships, job characteristics and worker satisfaction and performance. *Administrative Science Quarterly*, 26, 331-348.
- Brill, M. (1984). *Using office design to increase productivity* (Vol. 1). Buffalo, NY: Workspace Design and Productivity.
- Dean, L. M., Pugh, W. M., & Gunderson, E. K. E. (1975). Spatial and perceptual components of crowding: Effects on health and satisfaction. *Environment and Behavior*, 7, 225-236.
- Desor, J. A. (1972). Toward a psychological theory of crowding. *Journal of Personality and Social Psychology*, 21, 79-83.
- Epstein, Y. M., & Karlin, R. A. (1975). Effects of acute experimental crowding. *Journal of Applied Social Psychology*, 5, 34-53.
- Freedman, J. L. (1975). *Crowding and behavior*. San Francisco: Freeman.
- Fried, Y. (1985). *The validity of the job characteristics model: A review and meta-analysis*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.
- Hackman, J. R., & Lawler, E. E. (1971). Employee reactions to job characteristics [Monograph]. *Journal of Applied Psychology*, 55, 259-286.
- Hackman, J. R., & Oldham, G. R. (1975). Development of the Job Diagnostic Survey. *Journal of Applied Psychology*, 60, 159-170.
- Hayduk, L. A. (1983). Personal space: Where we now stand. *Psychological Bulletin*, 94, 293-335.
- Hom, P. W., Griffeth, R. W., & Sellaro, C. L. (1984). The validity of Mobley's (1977) model of employee turnover. *Organizational Behavior and Human Performance*, 34, 141-174.
- Jenkins, G. D., Glick, W. H., & Gupta, N. (1983). Job characteristics and employee responses. *Proceedings of the 43rd Annual Meeting of the Academy of Management*, 43, 164-168.
- Lange, H., Mueller, C. W., & Donnerstein, E. (1979). The effects of social, spatial, and interference density on performance and mood. *Journal of Social Psychology*, 109, 283-287.
- Mandel, D. R., Baron, R. M., & Fisher, J. D. (1980). Room utilization and dimensions of density: Effects of height and view. *Environment and Behavior*, 12, 308-319.
- Marshall, J. E., & Heslin, R. (1975). Boys and girls together: Sexual composition and the effects of density and group size on cohesiveness. *Journal of Personality and Social Psychology*, 31, 952-961.
- Mehrabian, A. (1977). A questionnaire measure of individual differences in stimulus screening and associated differences in arousability. *Environmental Psychology and Nonverbal Behavior*, 1, 89-103.

- Miller, H. E., Katerberg, R., & Hulin, C. L. (1979). Evaluation of the Mobley, Horner, and Hollingsworth model of employee turnover. *Journal of Applied Psychology*, 64, 509-517.
- Oldham, G. R., & Hackman, J. R. (1981). Relationships between organizational structure and employee reactions: Comparing alternative frameworks. *Administrative Science Quarterly*, 26, 66-83.
- Oldham, G. R., Hackman, J. R., & Pearce, J. L. (1976). Conditions under which employees respond positively to enriched work. *Journal of Applied Psychology*, 61, 395-403.
- Oldham, G. R., & Rotchford, N. L. (1983). Relationships between office characteristics and employee reactions: A study of the physical environment. *Administrative Science Quarterly*, 28, 542-556.
- Paulus, P. B. (1980). Crowding. In P. Paulus (Ed.), *Psychology of group influence* (pp. 245-289). Hillsdale, NJ: Erlbaum.
- Paulus, P. B., Annis, A. B., Seta, J. J., Schkade, J. K., & Matthews, R. W. (1976). Density does affect task performance. *Journal of Personality and Social Psychology*, 34, 248-253.
- Pierce, J. L. (1979). Employee affective responses to work unit structure and job design: A test of an intervening variable. *Journal of Management*, 5, 193-212.
- Pierce, J. L., Dunham, R. B., & Cummings, L. L. (1984). Sources of environmental structuring and participant responses. *Organizational Behavior and Human Performance*, 33, 214-242.
- Roberts, K. H., Hulin, C. L., & Rousseau, D. M. (1978). *Developing an interdisciplinary science of organizations*. San Francisco: Jossey-Bass.
- Rousseau, D. M. (1978). Measures of technology as predictors of employee attitudes. *Journal of Applied Psychology*, 63, 213-218.
- Saegert, S. (1978). High density environments: Their personal and social consequences. In A. Baum & Y. Epstein (Eds.), *Human responses to crowding* (pp. 257-281). Hillsdale, NJ: Erlbaum.
- Schiffenbauer, A. I., Brown, J. E., Perry, P. L., Shulack, L. K., & Zanzola, A. M. (1977). The relationship between density and crowding: Some architectural modifiers. *Environment and Behavior*, 9, 3-14.
- Schopler, J., & Stockdale, J. E. (1977). An interference analysis of crowding. *Environmental Psychology and Nonverbal Behavior*, 1, 81-88.
- Smith, S. W. (1978). Is there an optimum light level for office tasks? *Journal of the Illuminating Engineering Society*, 7, 255-258.
- Stone, E. F., & Porter, L. W. (1978). On the use of incumbent-supplied job-characteristics data. *Perceptual and Motor Skills*, 46, 751-758.
- Sundstrom, E. (1978). Crowding as a sequential process: Review of research on the effects of population density on humans. In A. Baum & Y. Epstein (Eds.), *Human responses to crowding* (pp. 31-116). Hillsdale, NJ: Erlbaum.
- Sundstrom, E. (1986). *Work places*. Cambridge, England: Cambridge University Press.
- Sundstrom, E., Burt, R. E., & Kamp, D. (1980). Privacy at work: Architectural correlates of job satisfaction and job performance. *Academy of Management Journal*, 23, 101-117.
- Sundstrom, E., Town, J. P., Brown, D. W., Forman, A., & McGee, C. (1982). Physical enclosure, type of job, and privacy in the office. *Environment and Behavior*, 14, 543-559.
- Vine, I. (1981). Crowding and stress: 1. Review of variables and theories. *Current Psychological Reviews*, 1, 305-324.
- Walden, T. A., & Forsyth, D. R. (1981). Close encounters of the stressful kind: Affective, physiological, and behavioral reactions to the experience of crowding. *Journal of Nonverbal Behavior*, 6, 46-64.
- Wineman, J. D. (1982). Office design and evaluation: An overview. *Environment and Behavior*, 14, 271-298.
- Worchel, S., & Teddlie, C. (1976). The experience of crowding: A two-factor theory. *Journal of Personality and Social Psychology*, 34, 30-40.

Received March 3, 1986

Revision received July 18, 1986 ■

Member Variation, Recognition of Expertise, and Group Performance

Robert Libby

Graduate School of Business Administration
University of Michigan

Ken T. Trotman

School of Accountancy
University of New South Wales, New South Wales, Australia

Ian Zimmer

Department of Commerce, University of Queensland, Queensland, Australia

The most effective method for aggregating the conflicting opinions of experts is a subject of active debate in the literature. Task differences are most often used to explain differing results among studies. Alternatively, we suggested that the characteristics of the interacting groups themselves determine whether they outperform or underperform their equivalent composites. Expert loan officers serving in ad hoc and practiced groups, on average, performed equally as well as did their composite and most influential individual. However, whether a particular group outperformed or underperformed its composite could be explained by variation in group members' performances and abilities to recognize differential expertise. These findings suggest the circumstances in which alternative social decision schemes are likely to be more effective. They also support the usefulness of conceptualizing group judgment as a weighted combination of the opinions of group members whereby the allocation of weights to members is the critical issue.

Individual versus group performance has been important in social psychology since the start of the century. The central issue in this research has been the extent to which the quality of group performance is above or below that of its members. More recently, however, there has been increased emphasis on the comparative performance of interacting groups and other alternative social decision schemes, particularly composite (or staticized) groups and best-member strategies. These analyses serve both a theoretical and an applied purpose. Comparisons of individual and group performance with these and other baseline models allow attribution of differences to a series of specific factors, increasing our understanding of group processes (e.g., Einhorn, Hogarth, & Klempner, 1977). They also provide a basis for selecting the most cost-effective decision scheme in different applied settings (e.g., Libby & Blashfield, 1978).

The results of studies on the comparative performance of interacting and composite groups have been somewhat conflicting. Most studies examined mean performance on a particular task or compared mean performance across a series of tasks. The considerable variation in performance among participant groups was usually treated as experimental error. As a result, differences in task characteristics were most commonly used to explain conflicting results (e.g., see Hill, 1982). No previous study has examined differences across groups within a specific context (task) to determine the characteristics of particular in-

teracting groups that do better or worse than their equivalent composite.

The purpose of the present study is twofold. First, group decision performance of expert loan officers working either in practiced or ad hoc groups is compared with various baseline models. This analysis follows Lorge, Fox, Davitz, and Brenner's (1958) suggestion of the need to extend the individual versus group performance literature from ad hoc groups solving trivial tasks to traditional or practiced groups solving important problems. Second, the impact of two group characteristics on the relative performance of interacting groups and their equivalent composites is analyzed. More specifically, we investigated the degree to which variation in group members' performances and the group's ability to recognize differential expertise determine relative performance.

Factors Affecting Comparative Performance

Research comparing interacting groups and composites (formed by taking the mean or majority of the individual judgments of the group members) has been summarized in a number of review articles with somewhat conflicting conclusions. Hackman and Morris (1975) concluded that research has shown that "for many tasks the pooled output of noninteracting individuals is better than that of an interacting group (p. 46)." Likewise, Fischer (1981), comparing interacting groups, composite groups, and the Delphi method on a probability forecast task, concluded that from a "practical viewpoint it makes little or no difference how one aggregates the conflicting opinion of experts" (p. 108). Alternatively, Rohrbaugh (1979) stated that "the empirical literature clearly points to the superior accuracy of judgments resulting from interacting groups in comparison to a baseline of staticized sets of individuals" (p. 75).

As we suggested earlier, most attempts to explain differences

We wish to thank Phil Yetton and Richard Hackman for their comments, and Jane Butt and Sarah Bonner for their assistance on this project.

Correspondence concerning this article should be addressed to Robert Libby, Graduate School of Business Administration, University of Michigan, Ann Arbor, Michigan 48109.

between studies have emphasized *task characteristics*. Even some of these explanations appear inconsistent. Hill (1982) concluded that for learning tasks and abstract problems, group performance was similar to that of the composite. Alternatively, in brainstorming and complex problem solving, the composite outperformed the interacting group. However, Yetton and Bottger (1982), using the National Air and Space Agency (NASA) moon problem, concluded that in multipart problems, which presumably involve brainstorming and complex problem solving, interacting groups outperform composite groups. This finding was consistent with previous studies using the same task (e.g., Holloman & Hendrick, 1970; Hall & Watson, 1970).

The key question that arises from these conflicting statements is still unanswered: Under what circumstances will an interacting group outperform or underperform a composite group? Hackman and Morris (1975) suggested an alternative emphasis; that is, the key issue in the comparative performance is the *group and its interaction process* through which new ideas, solutions, or efforts can be generated, which in turn can lead to process gain. Hill (1982) suggested two potential sources of this process gain, namely, member ability to learn and cognitive stimulation.

An alternative conception of the group and its interaction process was provided by Einhorn et al. (1977). They conceptualized the group judgment as a weighted combination of the judgments of the individual members and suggested that one crucial issue is the process used by the group to allocate weights to the opinions of different members. They developed four baseline models: (a) the random model, (b) the mean or composite model, (c) the best-member model, and (d) the proportional model. Through simulation, they showed that group size, the level of systematic bias, and the ability to recognize expertise affect the relative performance of these models. The third factor, the ability to recognize expertise, is of principle interest in the present study.

The importance of the ability to recognize expertise is shown in the Einhorn et al. (1977) analysis. First, as one would expect, the best-member model outperforms the composite model for all levels of bias. That is, if the group can select the best member with certainty, it outperforms the mean judgment of the individual group members. Second, the probabilities of selecting the best member in the proportional model are arbitrary, and as the probability of the group selecting the best member increases, the proportional model's performance improves and by definition equals that best-member model when the probability of selecting the best member equals one. This demonstrates how the group's ability to recognize relative expertise can determine whether it outperforms or underperforms its composite.

In addition, it can be inferred from Einhorn et al. (1977) that as bias increases, the ability to recognize differences in expertise becomes more important. For example, when standardized bias is zero, the mean model is substantially better than the proportional model, but inferior to the best-member model. As bias increases to one, the mean model is outperformed by the proportional model.

Yetton and Bottger (1982) provided additional support for the importance of the ability of group members to recognize relative expertise. Their results rejected error correction and

creativity as explanations for the superiority of interacting over composite groups. They suggested that interacting groups outperformed their equivalent composite groups because groups weighted individual inputs in proportion to an individual's expertise. They adopted the following model, which is an extension of the Einhorn et al. (1977) weighting scheme:

$$W_i = (N + 1 - i)^\omega / i^\omega,$$

where N = group size, i = the position of the person in the group with respect to ability on the task. As ω increases, increasing weight is given to the more proficient members. When ω is zero, this model equals a unit-weighted composite, and when ω equals infinity, the model provides a best-member strategy. Applying a $\omega = 2$ model, Yetton and Bottger found that the unequal weighted composite matched the performance of an interacting group. Yetton and Bottger concluded that "the use of differential ability provides a single and integrated explanation for previous disparate research findings" (p. 318) on the benefits of interacting groups compared to composites.

The evidence on the ability of group members to identify their best member has been mixed. Yetton and Bottger (1982) showed that groups can identify their best members with sufficient accuracy to enable the performance of an interacting group to equal that of its best member. Miner (1984), using a similar task, found that although groups appeared able to identify their better members, they identified their best member at little better than random chance. On an audit task, Trotman, Yetton, and Zimmer (1983) found that subjects could not differentiate relative expertise. However, these subjects were sampled from a restricted range of expertise.

The impact of the accuracy of expertise recognition on the relative performance of interactive and composite groups will be mediated by a second variable, the variation in the levels of expertise of the group members. At one extreme, if there is no variation in the members' judgments (they are exact replicates of one another), no scheme of differential weighting will improve the quality of group judgment. As the variation in judgmental accuracy among group members increases, so will the difference between the performance of the best member and the composite. If the interacting group can identify the more accurate members and weight their judgments more heavily in cases of increased variation, the benefit of this differential weighting will also increase.

The preceding discussion has suggested two characteristics of particular interacting groups that affect relative performance compared to composites: (a) ability to recognize relative expertise, and (b) variation in individual group members' performance. However, the degree to which these factors affect relative performance has not been specified because it depends on too many factors to be solved for analytically. This study concerns itself with the magnitude of the effects of these two factors on the relative performance of interactive groups and composites in a commercial lending context. In the experiment, participating loan officers predicted the future success or failure of 39 business firms under one of two conditions. In the first condition, experienced loan officers who had never worked together in a common work group participated. We will refer to these

as *ad hoc* groups. In the second condition, a second group of experienced loan officers participated as members of their established, or *practiced*, work groups. The first condition was designed to eliminate any potential bias due to hierarchical factors within an organization. The second captures any additional knowledge of relative expertise that may have been gathered through natural work relationships. Identical experimental protocols were used in each condition.

Method

Task

In the experiment, experienced Australian loan officers predicted whether 39 real but disguised companies would experience bankruptcy based on their financial profiles. This is a single-part judgment task where we would expect performance to be unaffected by factors such as the group's ability to generate new ideas or to bring differential expertise with problem subparts to bear. Similar tasks have been used in previous experiments by Libby and Blashfield (1978) and Zimmer (1981). In practice, such tasks are often performed in interacting groups.

The sample firms consisted of all of the companies listed on the Sydney Stock Exchange between January 1970 and December 1979 that were classified as land developers in the *Australian Stock Exchange Journal* and had survived for at least 5 years. Twelve of the sample firms failed. As this high failure frequency suggests, this industry is perceived by lenders to be quite risky.

Data for 3 years for five financial ratios were presented for each sample firm. These were (a) earnings (before interest and taxes) to total tangible assets, (b) cash flow (profits plus depreciation) to total liabilities, (c) current assets to current liabilities, (d) total liabilities to shareholders funds, and (e) retained earnings to total tangible assets. For the failed firms, these were the ratios for the 3 years prior to bankruptcy. For the nonfailed firms, the ratios were for 3 consecutive years randomly selected from the 10-year period. The environmental predictability of these ratios was tested by developing a discriminant analysis model for the year prior to failure or nonfailure and by testing the predictive accuracy of the models using the Lachenbruch cross-validation procedure. Using prior probabilities specified according to the correct environmental frequency of failure (12/27), environmental predictability was 33 out of 39 firms, a hit rate of 84.1%. The cases were presented in random order.

Subjects

The participants were 60 loan officers employed by Sydney merchant banks, trading banks, and finance companies. The participants in Condition 1, the *ad hoc* groups, were participating in one of two executive training courses at the University of New South Wales. These participants were randomly assigned to three-person groups. In Condition 2, the *practiced* groups, participation was arranged through an official in each of seven lending institutions that provided either 1 or 2 three-person groups. Unfortunately, one loan officer in Condition 2 misread the instructions and this group of three was dropped from the analysis, leaving 57 participants.

Procedure

The financial profiles, randomly bound into booklets, were initially distributed to each individual who completed the task independently. Participants were told how the sample of firms was selected (see earlier in Method section) and that their performance would be scored on an overall percentage-correct basis; that is, Type I or Type II errors have

equal costs. After all of the participants had completed the task on an individual basis, and prior to any feedback on their performance, they were allocated to a three-person interacting group. The participants then completed the same task as an interacting group. Although the booklets completed by individuals had been collected prior to allocation to groups, participants retained a copy of their individual answers. At the completion of the task, each group was asked to identify the person whom the group considered to be the most expert at this task.

Note that although the industry classification of the sample firms was provided, no information was given on the frequency of failure or nonfailure in the cases. As the cases contained the total population of the industry, knowledge of the true environmental frequency was a component of the expertise that the participants could bring to the task. Note that this procedure differs from previous studies using similar tasks in which unrepresentative frequencies have been used (see Libby & Blashfield, 1978, and Zimmer, 1981).

Participants in Condition 1 completed the tasks in class. The *practiced* groups completed the tasks in the offices of their respective banks. In all cases, at least one of the authors was present during administration.

Decision Schemes Considered

Six decision schemes are considered in this study: the average individual, composite, interacting group, actual best member, subjective best member, and the most influential member. The actual best member was identified by the researchers *ex post facto*, and the subjective best member was the member so selected by the group after completion of the task. The most influential member was determined by counting the number of times each member's opinion was followed by the group. This method is consistent with that used by Bottger (1984).

Measurement of Variables

The performance or accuracy of the alternative social decision schemes was determined by counting the number of cases out of 39 that were correctly classified. Variation in expertise was measured by the standard deviation of the accuracy scores of the individual group members. Ability to recognize expertise was measured in two ways. First, the position of the subjective best member in the group with respect to accuracy was measured—*most accurate* (1), *second most accurate* (2), or *least accurate* (3). Although we have no data on the actual process the groups used to select the subjective best member, it is likely that this first measure is subject to considerable error due to social pressures relating to hierarchical position or experience affecting the nomination of the best member, a lack of insight into the group weighting process (see Slovic & Lichtenstein, 1971), or both. As an alternative, this study also measures ability to recognize expertise by the accuracy ranking of the most influential member, that is, whether he or she was the *most* (1), *second most* (2), or *least* (3) accurate member. In both cases, the higher the rank, the more accurate is the group's judgment of relative expertise.

Results

As shown in Table 1 there are no significant differences (*t* tests, $p > .05$) between the performance of the participants from *ad hoc* and *practiced* groups, regardless of the decision scheme used. This suggests that the impact of more accurate knowledge of relative expertise in the *practiced* groups (if any) was offset by the effect of social pressures. Consequently, the results from Conditions 1 and 2 are combined in all further analysis. To allow comparison with prior studies, the accuracy of the alterna-

Table 1
Comparison of Ad Hoc and Practiced Groups

Decision scheme	Ad hoc		Practiced		All groups	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Average individual	29.2	2.06	30.0	2.01	29.6	2.02
Actual best member	31.7	1.49	32.7	1.72	32.2	1.65
Composite group	31.6	2.07	31.2	3.15	31.4	2.57
Interacting group	31.0	1.70	30.9	2.71	30.9	2.17
Subjective best member	29.7	3.09	30.1	2.67	29.9	2.83
Most influential	30.7	1.89	30.9	3.02	30.8	2.42

tive decision schemes are first compared. Then the primary analysis of the factors that affect comparative performance is presented. Finally, some additional results relating to the ability of interacting groups to recognize expertise are reported.

Comparative Performance of Decision Schemes

To allow comparison with prior studies, the accuracies of the alternative decision schemes were compared using a Newman-Keuls test and are presented in Table 2. The difference between the performance of the individuals and groups can be divided into two components. First, the difference between individual and composite performance can be attributed to reduction of error variation in the individual judgments (an assembly or diversification effect; see, e.g., Yetton & Bottger, 1982). It can be shown mathematically that the composite will outperform the average individual if the correlation between responses of any pair of individuals is less than one (see, e.g., Dawes, 1970, and Libby & Blashfield, 1978). Second, any difference between the interacting group and the equal weighted composite results from marginal process gain or loss (an interaction effect). Consistent with most previous studies (see reviews by Hill, 1982, and Shaw, 1981) the composites and interacting groups outperformed the average individual ($p < .05$). However, there was no significant difference in performance between the composites and interacting groups ($p > .10$). This suggests significant assembly gains from diversification of error, but no marginal process gains from interaction.

Comparisons with the actual best member and most influential member are used to provide insight into the sources of and potential for certain types of process gain or loss. The fact that the most influential group member was significantly more accurate than the average individual implies some degree of accuracy in the group members' judgments of relative expertise. Yet, the significantly superior performance of the actual best member ($p < .05$) over the most influential may indicate the potential for further gains if the most accurate member can be more reliably determined.¹ Given that there was no significant difference in the performance of the composites and most influential individuals,² the potential gain from recognition of expertise in the current study was approximately equal to the potential gain from the assembly effect. However, the surprising finding that the performance of the most influential member was virtually identical to that of the interacting groups indicates that the in-

teraction process does not combine these two benefits with cognitive stimulation and learning to produce any marginal effect over and above that of the individual effect of expertise identification or assembly.

Factors Affecting Group Performance

Although, on average, there was no significant performance difference between the interacting groups and composites, there was considerable variation in this relationship across the participating groups. Table 3 provides a correlation matrix of the group characteristics expected to affect the relative performance of specific interactive groups and their equivalent composites: group minus composite performance and the two measures of ability to recognize expertise and variation in expertise. A full correlation matrix of all variables in Tables 1 and 3, plus "composite less average individual performance" and "interacting group less average individual performance" is included in the Appendix (Table A-1). As expected, the difference between group and composite performance was significantly correlated with the variation in the accuracy of the group members' judgments. In addition, the higher the position (in terms of accuracy ranking) of the most influential member, the higher was interacting group minus composite performance. This suggests that the ability to *actually* weight judgments based on relative expertise is a significant determinant of group performance. Further, these two determinants of group performance themselves were not significantly correlated.

Although the accuracy ranking of the subjective expert was significantly correlated with the accuracy ranking of the most influential member, it was *not* correlated with group minus composite performance. These two results taken together suggest significant error in the subjective measure as discussed in an earlier section. When the same factors in Table 3 were correlated with group less individual performance, very similar coefficients resulted (.48 for variation and $-.52$ and $.11$ for the two measures of ability to recognize expertise).

¹ Ex post facto selection of the best member in the current study may overstate the practical benefit of this improvement.

² The most influential member was the same person who agreed most with the composite in 12 of 19 groups. The correlation between the performance of the most influential members and the members who agreed most with the composite was .29.

Table 2
Comparative Performance of Alternative Social Decision Schemes

Decision scheme	Difference between means					
	1	2	3	4	5	6
1. Individual mean	—	0.31	1.21*	1.37**	1.84**	2.73**
2. Subjective best member		—	0.89*	1.05	1.53**	2.32**
3. Most influential			—	0.16	0.63	1.42**
4. Interacting group				—	0.47	1.26**
5. Composite group					—	.79
6. Actual best member						—

* $p < .10$. ** $p < .05$.

To assess the joint contribution to group performance of ability to recognize expertise and variation in expertise, a regression model was constructed with group minus composite performance as the dependent variable. The two independent variables were the variation in individual performance (Factor 1) and the position of the most influential member (Factor 2). The squared multiple correlation coefficient (R^2) was .5254, and both factors were significant (Factor 1, $t = 3.24$, $p = .005$; Factor 2, $t = -2.36$, $p = .032$). As expected, adding the second measure of ability to identify expertise to the model did not increase predictability.

Further Analysis

As was noted earlier, there has been conflicting evidence on the ability of an interacting group to recognize its best member (see Miner, 1984; Yetton & Bottger, 1982). Results here showed that interacting groups were capable of selecting the best member at better than random chance ($p = .06$, binomial test). Further, there was no difference in the performance rankings of the person selected as the best member and the person who was most influential in the group. This appears to conflict with our previous suggestions that there was more error in the subjective selections of the best member than in determination of the most influential individual. However, further analysis suggests that there are indeed differences. Their source is the magnitude of the errors in ranking. There were four instances where the most influential individual outperformed the subjective expert. In these cases, on average, the difference in performance was 14%. At the same time, in the three instances where the subjective expert outperformed the most influential member, the difference was only 4%.

Discussion

In an attempt to explain between-group differences and conflicting findings from prior research, this study addressed the factors that may affect the comparative performance of interacting groups and composites. Unlike prior studies in which task differences have been the primary explanatory variables, this factor was held constant in the current study, and attention was focused on two characteristics of the groups themselves. The current results suggest that these two factors, the ability to rec-

ognize expertise and the variation in individual performance, are major determinants of group versus composite performance. They explained fully 53% of the variation in this difference.

The importance of the first factor is consistent with the analytical work of Einhorn et al. (1977), and supports the theme of differential expertise and information aggregation developed by Yetton and Bottger (1982). This study demonstrated that as the ability to recognize expertise increases, the difference between interacting group and composite performance increases.

The relation between the variation in expertise and interacting group versus composite performance is also consistent with Yetton and Bottger's (1982) analysis. If differences in performance cannot be explained by error correction or creativity and there is no variation in expertise, then the best member can perform no better than a composite. Consequently, an interacting group cannot outperform a composite. Consistent with this analysis, the level of variation in individual judgments was significantly related to the comparative performance of interacting and composite groups.

Yetton and Bottger (1982) suggested that the question of the ability of a group to identify its best member has been a neglected theme in the small-group literature. They found that groups could recognize their best member with sufficient expertise for the accuracy of the selected best member to equal the

Table 3
Correlation of Group Characteristics

Group characteristic	1	2	3	4
1. Group less composite performance	—			
2. Variation in individual performance	.60	—		
3. Position of subjective expert	.08	.26	—	
4. Position of most influential member	-.46	-.10	.63	—

performance of an interacting group. However, Miner (1984) found for his IGB (individual judgment, followed by group judgment, followed by selection of the best member) strategy, which was similar to that used in the current study, that subjects only selected the best member in 22% of the cases. This was slightly worse than random (random probability equaled 26%). In addition, inconsistent with Yetton and Bottger, Miner found that the interacting group performance was significantly better than the performance of the subjective best member. The current results suggest that the difference between the studies may be caused by error in measuring the subjective best member.

Miner (1984) found that an interacting group equaled the performance of the actual best member. If, like Yetton and Bottger (1982), we reject error correction explanations, this could only occur if the groups were actually placing most weight on their best (or possibly better) members. If groups were indeed weighting their best member, but the knowledge of expertise implicit in these weightings was not conveyed in their verbal subjective assessments, then the verbalization process itself must be a source of significant error. This error in measurement must be higher in Miner than in Yetton and Bottger, which may be due to social factors caused by the method of forming groups. Although Yetton and Bottger randomly allocated subjects to groups, Miner allowed individuals who had previous interactions to form their own groups.

The present article provides additional evidence on this best-member theme. Loan officer groups could select their best member with better than random probability, but these selections were just barely accurate enough to allow the performance of this subjective best-member strategy to equal the performance of an interacting group. This places the accuracy of the current subjects' subjective estimates between those of the prior studies. We suggest that differences among the studies are primarily a function of the social setting in which the measures were made. This provides further support for Bottger's (1984) results and analysis of the significance of distinguishing between actual and subjective measures of influence in modeling group behavior.

However, the interacting groups in the present study did not appear able to actually *weight* expertise (as measured by the position of the most influential member) as well as in either the Miner (1984) or Yetton and Bottger (1982) studies. For example, in contrast to Miner's study, the interacting group was significantly outperformed by the actual best member. In addition, we found that composite groups outperformed the selected best member, which is inconsistent with Yetton and Bottger. These differences may have been task related. The bank loan officers would be expected to have more similar backgrounds than the managers and MBA students who completed the NASA moon problem or the winter survival problem. For example, all loan officers have been trained to use accounting ratios, but knowledge of the moon and survival tactics in general would be expected to vary considerably between, say, science and business graduates. This variance would be expected to affect the ability to differentiate expertise and, therefore, the weighting of expertise.

From a practical standpoint, the fact that interacting groups, composites, and most influential individuals all performed

equally suggests the two dominant strategies. Either the meeting of the group members could be eliminated or, after the most influential member is determined, the other two members could be released. More accurate selection of the best member also has the potential to produce additional gains. Findings that practiced groups were no more able to recognize expertise than ad hoc groups of strangers indicate that this may be difficult to improve in practice. In practiced groups, it is possible that the impact of more accurate knowledge about expertise is offset by negative effects of greater social pressure. However, other benefits of interactive groups related to gaining commitment, the training of novice decision makers, and the reduction of individual information load must also be considered in making this choice.

In conclusion, note that both ability to recognize expertise and variations in individual expertise are affected by the type and complexity of the task. Recall that our task involved evaluation of a limited data set within a very narrow context (a single industry). In multipart tasks, in which expertise with different subparts of the problem or information set is likely to vary more, or in situations in which judgments must be made across a wide variety of contexts, performance variation and ability to recognize expertise may both be greater. In such situations, the relative performance of the different social decision schemes may differ. Thus, although the task was kept constant in this study, our analysis suggests that the type of task will have an important effect on comparative performance. However, the more important conclusion that can be drawn is that these task effects can be effectively explained by their impact on variation and ability to recognize expertise.

References

- Bottger, P. (1984). Expertise and air time as bases of actual and perceived influence in problem-solving groups. *Journal of Applied Psychology*, 69, 214-221.
- Dawes, R. M. (1970). *An inequality concerning correlation of composites vs. composites of correlations* (Oregon Research Institute Methodological Note No. 1). Eugene: University of Oregon.
- Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin*, 84, 158-172.
- Fischer, G. W. (1981). When oracles fail—comparison of four procedures for aggregating subjective probability forecasts. *Organizational Behavior and Human Performance*, 28, 96-110.
- Hackman, J. R., & Morris, C. G. (1975). Group tasks, group interaction process and group performance effectiveness: A review and partial integration. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 8, pp. 47-99). New York: Academic Press.
- Hall, J., & Watson, W. H. (1970). The effects of a normative intervention on group decision-making performance. *Human Relations*, 23, 299-317.
- Hill, G. W. (1982). Group versus individual performance: Are $N + 1$ heads better than one? *Psychological Bulletin*, 91, 517-539.
- Holloman, C. R., & Hendrick, H. W. (1970). Individual versus group effectiveness in solving factual and nonfactual problems. *Proceedings of the 78th Annual Convention of the American Psychological Association*, 5, 247-254.
- Libby, R., & Blashfield, R. K. (1978). Performance of a composite as a function of the number of judges. *Organizational Behavior and Human Performance*, 21, 121-129.

Lorge, I., Fox, D., Davitz, J., & Brenner, M. (1958). A survey of studies contrasting the quality of group performance and individual performance. *Psychological Bulletin*, 55, 337-371.

Miner, F. C. (1984). Group versus individual decision making: An investigation of performance measures, decision strategies, and process losses/gains. *Organizational Behavior and Human Performance*, 33, 112-124.

Rohrbaugh, J. (1979). Improving the quality of group judgment: Social judgment analysis and the Delphi technique. *Organizational Behavior and Human Performance*, 24, 73-92.

Shaw, M. E. (1981). *Group dynamics*. New York: McGraw Hill.

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, 6, 649-744.

Trotman, K. T., Yetton, P. W., & Zimmer, I. R. (1983). Individual and group judgments of internal control systems. *Journal of Accounting Research*, 21, 286-292.

Yetton, P. W., & Bottger, P. C. (1982). Individual versus group problem solving: An empirical test of a best-member strategy. *Organizational Behavior and Human Performance*, 29, 307-321.

Zimmer, I. R. (1981). A comparison of the prediction accuracy of loan officers and their linear-additive models. *Organizational Behavior and Human Performance*, 27, 69-74.

Appendix

Table A-1
Complete Correlation Matrix

Variable	1	2	3	4	5	6	7	8	9	10	11
1.	—										
2.	.62	—									
3.	.85	.51	—								
4.	.68	.64	.81	—							
5.	.65	.18	.50	.33	—						
6.	.37	.48	.37	.69	.50	—					
7.	-.66	.14	-.56	-.24	-.59	.03	—				
8.	-.02	.25	.01	.07	-.73	-.36	.26	—			
9.	.20	.15	.09	-.22	-.29	-.64	-.10	.63	—		
10.	.13	.06	.63	.53	-.01	.16	-.07	.05	-.14	—	
11.	-.32	.08	.02	.48	-.34	.44	.48	.11	-.52	.53	—
12.	-.48	.04	-.54	.06	-.38	.35	.60	.08	-.46	-.30	.65

Note. 1 = average individual performance; 2 = actual best member performance; 3 = composite group performance; 4 = interacting group performance; 5 = subjective best member performance; 6 = most influential member performance; 7 = variation in individual performance; 8 = position of subjective expert; 9 = position of most influential member; 10 = composite less average individual performance; 11 = interacting group less average individual performance; 12 = interacting group less composite performance.

Received April 1, 1986
Revision received July 31, 1986 ■

Routinization of Mental Training in Organizations: Effects on Performance and Well-Being

Gerry Larsson

Division of the Behavioural Sciences, The Swedish National Defence Research Institute, Karlstad, Sweden

The purposes of this study were twofold. The first was to improve performance in an organizational setting (military) with the help of the following mental-training techniques: relaxation, meditation, and imagery rehearsal—alone or combined. Second was to routinize the process using the organization's staff as coordinators, minimizing teacher training, preparation, cost, and time. A training program was developed within these guidelines. An experimental group of 214 Swedish conscripts and cadets (men, 19–21 years old) followed the training program for approximately 8 months. The performance of this group was significantly better than the control group on actual task examinations and mental tests. No effects from the training program were found on physical and mental well-being. Most goals concerning the routinization of the training were reached. Necessary conditions for and possible benefits from large-scale applications of mental training in organizational contexts are discussed.

Autogenic training, biofeedback, hypnosis, meditation, progressive relaxation, and other methods have been used as aids against stress in various fields—psychotherapy and sports in particular. Lazarus (1966), in his classical work on stress and coping, defined the acute stress reaction as being composed of physiological, motor behavioral, cognitive, and emotional elements. The physiological level involves a multitude of changes governed by the autonomic nervous system. On the motor-behavior level, significant arousal seems to activate not only the appropriate muscles for a given task but also their antagonists, which ideally should be relaxed (Chaney & Andreasen, 1973). The results are wasted energy, jerky movements, and, in extreme cases, paralysis. On the cognitive level, too much arousal can lead to increased conscious attention to one's own process of performance that may result in a performance decrement (Baumeister, 1984). The emotional aspect of the acute stress reaction includes the following unpleasant emotions: anger, anxiety, depression, fear, guilt, and shame.

Various performance decrements and disturbances in well-being are frequently reported as long-term effects of stress. In the work force, Cox (1978) claimed that reductions in productivity and job satisfaction may be attributed to external and internal stressors. It is therefore little wonder that relaxation and meditation techniques have attracted considerable interest as a coping technique for stress.

Most research on the effects of relaxation and meditation has been carried out on the physiological and/or psychological level(s), studying the variables of performance and health. A summary of previous research findings follows. This summary is limited to findings from studies of "normal" nonclinical subjects; pre-post assessments and control groups were used. No

differentiation will be made between different techniques of relaxation and meditation. Advocates of the different schools frequently claim unique (and remarkable) results. However, Setterlind (1983) in an extensive study found that different techniques lead to similar physiological and psychological (and less remarkable) results. No method seems to be universally superior, and choice of technique should ideally depend on the individual. A guide to method selection, based on the individual's physical and mental levels of tension, is presented by Davidson and Schwartz (1976).

On the physiological level, relaxation and meditation have been found to lead to (a) a reduction of oxygen consumption, rate of respiration, heart rate, blood pressure, muscular tension, and concentration of lactic acid in the blood, and (b) an increase in alpha waves, skin resistance, and skin temperature (Benson, 1975; Setterlind, 1983).

On the motor-behavior level, most research on the effects of relaxation and meditation has been carried out on students and athletes. In summarizing these studies, some demonstrate performance improvements (Appelle & Oswald, 1974; Decaria, 1977; DeMers, 1979/1980; DeWitt, 1980; Garver, 1977; Hickman, Murphy, & Spino, 1977; Nideffer & Deckner, 1970; Pardine, Dytell, & Napoli, 1981; Scott & Pellicioni, 1982; Suinn, 1980; Weinberg, 1982), and others show no results (Arnold, 1969/1970; Chaney & Andreasen, 1973; Gordin, 1981; Klisch, 1980/1981; Nelson, 1980; Williams & Herbert, 1976; Williams & Vickerman, 1976). Negative results were not shown. According to Setterlind (1983), the studies showing no results seem to be better designed than those showing positive results. Most studies suffer from small samples. In my opinion, the validity of the results achieved in laboratory settings on real-life motor tasks is questionable, as is the reliability and validity of the criteria measurements used in most of the real-life studies (e.g., basketball dribbling and dart throwing). In conclusion, large-scale studies of actual motor task behavior in organizational settings with adequate performance measurements are lacking.

I appreciate the comments of C'Anne Cook-Steffy on an earlier draft of this article.

Correspondence concerning the article should be addressed to Gerry Larsson, FOA 55, Karolinen, 651 80 Karlstad, Sweden.

On the cognitive level, in a computer search of the literature, few studies that satisfy the criteria stated here were revealed. Concerning basic mental abilities, Chaney and Andreasen (1972) found that progressive relaxation led to significant improvement on a memorization test (recall of random numbers). Pelletier (1974) found that meditation led to increased field independence. Meditators claimed to experience an increased alertness and awareness of their surroundings. These claims are consistent with the findings of Anand, Chhina, and Singh (1961), who showed that meditators appeared to lack habituation to alpha blocking. In an applied context, Barabasz (1980) showed that hypnosis led to improved performance on a simulated radar detection task. As with the studies on motor performance, many of the experiments are based on comparatively small samples of subjects who have practiced relaxation or meditation for a short time.

On the emotional level, which includes the general aspects of well-being, the effects of relaxation and meditation have been dealt with in several good studies. However, a problem with these studies, in relation to the criteria stated, is that many focus on people who suffer from various kinds of anxiety. The effect of relaxation and meditation on everyday life, mood changes, and on the well-being of adequately functioning people, has not received much scientific attention. However, there are some exceptions; the most notable of which is a large-scale study carried out by the U.S. Navy Air Schools during the Second World War (Neufeld, 1951). In this study, about 15,000 cadets practiced progressive relaxation three times per week for 10 weeks. Reduced injuries and improved sleep were the main results. Using Benson's (1975) meditation procedure in the work context, Peters, Benson, and Porter (1977) found positive results (self-reports) on performance and well-being. Using progressive relaxation and meditation on a working population, Carrington et al. (1980) found positive results (self-reports) on stress reduction and well-being. After introducing fixed rest periods in a work setting, Bhatia and Murrell (1969) reported improvements in production and increased job satisfaction. Setterlind (1983), in a study with a large sample of school children, found positive results from progressive relaxation and meditation on well-being. Characteristic of all the aforementioned studies (except for the Neufeld study, where no pretraining assessment of anxiety is presented) and exemplified by Setterlind's study, the individuals who showed high levels of anxiety at the onset of the training benefited the most from it. No significant changes on well-being variables were found among the low-anxiety subjects. Noteworthy of these studies is their relatively short training period—2 to 3 months.

In behavior therapy and sport psychology there is often a combination of relaxation or meditation and imagery. During imagery, one fantasizes or "experiences" oneself performing the task (Browne & Mahoney, 1984; Wolpe, 1958). During relaxation or meditation (after a period of practice), one is said to reach an altered state of consciousness (Korn & Johnson, 1983; Setterlind, 1983). This state is physiologically characterized by an increased level of alpha activity (Brown, 1970; Danskin, 1981; Girdano & Everly, 1979; Kamiya, 1969). In this state, as well as in the dream state, the right hemisphere seems to be more active than the left (Bakan, 1969; Blumberger, 1978; Galin, 1974; Rubenzer, 1979). The right hemisphere also ap-

pears to be superior on tasks involving spatioperceptual ability (e.g., visual imagery; Springer & Deutsch, 1982). Thus, imagery is claimed to be more vivid and realistic in the alpha state than during the state of ordinary wakefulness (beta).

There are several theories (not necessarily conflicting) concerning the physiological changes during imagery and why it may lead to actual performance improvement (Bakan, 1980; Richardson, 1983; Sage, 1971). A neuromuscular theory hypothesizes that imagery activates neuromuscular units involved in a specific task, although no actual movement can be observed (Brown, 1980; Garver, 1977; Sage, 1971; Shaw, 1940). Another example is a neurophysiological theory that hypothesizes an activation of specific neural components in the brain involved in the direction of the movements being imagined while the actual motor component is not involved (Garver, 1977; Sage, 1971).

Imagery can be experienced in two ways. One can either "see" oneself as an observer, or one can experience the process from within, as is the case in reality. The first method, external imagery, relates primarily to perception. The second method, internal imagery, involves all senses and relies on memory and integrated knowledge (Cook, 1984). An exploratory study by Mahoney and Avenier (1977) on elite gymnasts suggests that internal imagery may be more efficient.

Studies of imagery rehearsal training and relaxation or meditation in sport contexts (Clark, 1960; Decaria, 1977) indicate that a reasonable level of actual skill is necessary before one can expect performance benefits. No improvement is frequently reported among novices. This seems reasonable inasmuch as the use of imagery requires that the individual has internalized a well-structured mental representation of the ideal behavior. On most complex tasks this is difficult to achieve at the novice stage.

In the literature, two areas can be cited where systematic studies are lacking. The first area deals with the possibility of achieving actual performance improvements in organizational settings by using relaxation and meditation techniques alone or combined with imagery. The second area that requires further study questions the routinization of this training; that is, training should be led by ordinary organization personnel (supervisors or teachers) and not by specialists (psychologists). Routinization also implies a longer training period than was the case in the studies I have referred to. This article focuses on these two points. It was predicted that relaxation training or meditation alone or combined with imagery rehearsal training, led by ordinary organizational staff over at least a 6-month period would lead to (a) improvements in actual task performance, (b) improvements of psychological functions mainly conducted by the right hemisphere (e.g., visiospatial ability)¹, and (c) increased well-being. For the sake of parsimony, and in agreement with the current praxis in Sweden, relaxation training or meditation alone or combined with imagery rehearsal training, will henceforth be labeled *mental training*. Before describing the method and results of the present study, I will present a brief

¹ The study of effects of mental training on general psychological functions, for example, visiospatial ability, is reported in detail elsewhere (Larsson, Starrin, Olofsson, & Bäck, 1986). In this article, only a summary of these results is given.

summary of the pedagogical model used for teaching and learning mental training.

A Model for Teaching and Learning Mental Training

The pedagogical model for teaching and learning mental training was guided by certain goals. The goals of the organization were the following: The mental training program should be time efficient, cost efficient, and should interfere as little as possible with the everyday tasks. The same goals were to apply to the education of the organizational staff (platoon officers) who were to lead the training. The goals of the program participants (cadets and conscripts) were the following: The training program should be structured to allow each participant to discover the most suitable technique or combination of techniques, to learn the chosen method thoroughly and quickly, and to be able to use it independently, thereby increasing his general action competence.

The model for teaching and learning relaxation or meditation (alone or combined with imagery) was designed to accommodate six phases.

1. Training of teachers by a specialist. The teachers receive a theoretical orientation and some guidelines on the teaching of mental training. They also practice the same program the participants will follow later.

2. Introduction conducted by the teacher for the participants. The main objectives of the theory are summarized, and the training program is described.

3. Basic relaxation training is established at intervals of four times per week for 4 consecutive weeks. The teacher plays tape-recorded programs beginning with progressive relaxation and then moving on to more passive programs where one merely thinks through the various parts of the body without first tensing the muscles.

4. Deep-relaxation training once or twice a week for about 25 weeks. The teacher introduces a simple meditation technique (Benson's, 1975, "one" meditation). The participants may then choose between progressive relaxation, a tape-recorded program with flute music, or meditation. However, they are encouraged to practice by themselves without a tape recorder.

5. Applying relaxation training in stressful situations. The relaxation reaction is conditioned to a stimulus—tensing the muscles of one's left hand tightly while holding one's breath for several seconds—then release. This conditions the stimulus to work as a *trigger*, a special signal, which rapidly produces relaxation. Participants are encouraged to practice the trigger in various stressful situations.

6. Imagery rehearsal training of the task at hand. This is introduced by the teacher when the participants have acquired a reasonable level of skill at their task and have finished the basic relaxation training program. In principle, the program has the following character: relax—imagine—wake up.

A final requirement of the pedagogical model is that a psychotherapeutically trained person supervise the program teachers and be available for advice or active help. It is possible that relaxation and meditation may activate repressed unpleasant memories and subsequently cause increased levels of anxiety (Østbye-Sundsvold & Vaglum, 1984).

Method

The present study was carried out on military conscripts and cadets in Sweden. Military service is compulsory for all men, and most do their basic training (about 8 months) at the age of 19 or 20. A limited number of noncommissioned officers go to cadet school (10 weeks) immediately following basic training. Conscripts, cadets, and officers from the anti-aircraft artillery and the signal troops constituted the participants in this study.

Subjects

In a system as highly planned and organized as the military, it is difficult to achieve randomization. In this experiment, however, it was possible to randomize treatment conditions among groups (i.e., platoons). Due to the lack of comparable platoons, randomization among groups was limited. No subdivisions within platoons were practical. The experimental group consisted of 214 conscripts and cadets (five platoons with conscripts and two platoons with cadets). The control group (conventionally trained without mental training) consisted of the following: 116 conscripts (four platoons) on task performance, 154 conscripts and cadets (two platoons with conscripts and three platoons with cadets) on mental tests, and 32 conscripts (one platoon) on the well-being assessment.

All of the subjects were men, approximately 19 to 21 years old. The platoons were stationed at four different locations in Sweden. None of the subjects knew they would participate in this study, prior to beginning their military service. All conscript subjects were complete novices at the task they were to learn; the cadets had learned their tasks and had received some leadership training during basic training. The main task for the anti-aircraft artillery conscripts was a perceptual-motor task—manual operation of a laser-guided anti-aircraft missile. The main task for the signal troops conscripts was a cognitive task—receiving Morse code signals. The main task for the anti-aircraft artillery cadets was a social competence task—development of general leadership competence.

Training of Teachers

The platoon officers were taught mental training in theory and practice by the author (Phase 1). The training consisted of two lessons (4 hr per lesson), with an intermediate period of 4 weeks, during which relaxation was practiced individually. None of the officers had any formal education in psychology, except for some minor courses in social psychology and leadership. The platoon officers did not volunteer to take part in the study; they were simply ordered to carry it out.

Training Aids

The platoon officers were given the following articles: a popular report on stress and relaxation (Setterlind & Larsson, 1983), a well-known report on relaxation, imagery, and performance (Larsson & Setterlind, 1983), and a book on stress and psychosomatic health (Larsson, Setterlind, & Steffner-Starrin, 1985). Four tape-recorded programs (Setterlind & Larsson, 1983) were used for basic relaxation training and conditioning to the trigger. These programs were inspired by Uneståhl's (1979) Swedish interpretation of progressive relaxation. Two additional recordings (Larsson, 1984), flute music and "one" meditation, were used for deep-relaxation training.

Four tape-recorded programs (Larsson & Bäck, 1985) were used for imagery training. In each of these programs, the subject was first directed to visualize the general scenery, including himself (external imagery), then he was asked to imagine and experience the task behavior from within (internal imagery).

Procedure

The platoon officers (3 or 4 per platoon) conducted the mental training according to the model presented above. All experimental platoons received relaxation training and meditation (Phases 3 to 5). One experimental platoon with conscripts from the anti-aircraft artillery ($n = 32$) participated in the imagery rehearsal training (Phase 6).

The conscripts and cadets were trained in relaxation and meditation at the convenience of their platoon officers. Training was regular but flexible, and not, for example, systematically carried out a short time before stressful examinations. For the conscript subjects from the anti-aircraft artillery, relaxation training was not to interfere with their main task—learning to operate the laser-guided anti-aircraft missile. However, imagery training for one designated platoon was intentionally carried out during actual task training time, thereby limiting the actual task training time for these experimental subjects.

Signal troop conscripts received half of their relaxation training and meditation during actual task training time, similarly reducing their actual task training time.

The following comparisons were made between the experimental and control groups: (a) physical and mental status 1 year before the military basic training (results from the compulsory enlistment examination were used), (b) questionnaire and interview ratings (made at three different times by the conscripts and once by the cadets) on how they perceived their military service and the training program, and (c) performance on a selected number of main task examinations during the basic training.

The selection of main task examinations was made in advance by military experts following criteria of selection based on stressfulness. The examination data consisted of reliable numerical figures. The conscripts from the anti-aircraft artillery were examined by a training simulator that electronically calculated the error for each shot (height and direction). The conscripts from the signal troops were evaluated by the number of errors made while receiving Morse code signals.

No objective comparison data existed for one platoon of conscripts from the signal troops ($n = 18$) or for the cadets from the anti-aircraft artillery ($n = 80$). Thus, for these two groups the effects of mental training on task performance could only be assessed by self-ratings and by ratings made by their platoon officers. All of the platoon officers (but one) had previous experience training similar platoons.

Another comparison was made on performance on psychological tests that measured left-hemisphere functions (verbal and numerical-sequential tests) and functions of the right hemisphere (visiospatial tests).

The final comparison was a pre-post comparison on a 47-item questionnaire designed to measure well-being. The questions measured self-image (ideal, real, and social self), general as well as situation-specific anxiety, psychosomatic health, smoking, drinking, physical exercise habits, previous experience of mental training, and perceived need of mental training.

Results

Pretraining Assessments

No significant differences were found between the experimental and control groups at the enlistment examination on either physical or psychological variables. In general, the subjects were above average physically and mentally. For example, the level of intelligence for conscripts from the anti-aircraft artillery averaged about 5.0 (stanine values); the conscripts from the signal troops and the cadets had an average of about 7.0. The intelligence test at the enlistment examination consists of four subtests designed to measure reasoning, verbal ability, spa-

tial ability, and technical understanding. Scores are transformed to stanine values; the comparison group was composed of Swedish male 18-year-olds. The similarity between the experimental and control groups was predictable because selection requirements for Swedish conscripts are very detailed—therefore, those in a given position generally have similar physical and psychological characteristics.

Before the start of mental training, the participants in the experimental group were asked about previous experiences with this kind of training. Approximately 31% had tried relaxation training on at least one occasion. Most of these had tried it either at school or in sport contexts. No one had practiced mental training regularly.

Training Conditions

Mental training. The mental training program was carried out with a slightly lower training frequency than planned. Basic relaxation training (Phase 3) consisted of about 12 sessions instead of the planned 16. Deep-relaxation training (Phase 4) consisted of about 25 sessions instead of the planned 35 to 40. Applied relaxation training in stressful situations with the help of a trigger (Phase 5) was taught to six platoons ($n = 182$). All of the subjects took part in about 35 to 40 relaxation or meditation sessions. Imagery training (Phase 6) was carried out according to plan on one of the platoons ($n = 32$) from the anti-aircraft artillery. Each of the conscripts in this platoon took part in about 10 to 15 imagery training sessions. The mental training occupied about 1% of total military training time.

Actual task training. Ratings made by the platoon, company, and battalion officers indicated that all experimental and control platoons involved in the study had comparable actual task-training conditions. Because some of the mental training took place during actual task-training time (see the Method section), the subjects in the experimental platoons received about 3–5% less actual task training time than the subjects in the control platoons.

General military training. In summary, it can be concluded that the conscripts and cadets of the experimental and control platoons perceived their military training similarly, and their platoon officers were rated as equally good by company and battalion officers. No obvious signs of a Hawthorne effect were directly observed or noticed in the interviews. In light of what is experimentally desirable, the data were ideal. However, the data also indicate that the mental training program most likely did not have any significant effect on how the conscripts and cadets perceived their military training.

Mental Training Experience

Collection of self-report data concerning how the conscript subjects perceived the mental training program occurred during the 5th, the 15th, and the 25th week of the program. The cadets were interviewed and questioned once during the 5th week of training. Measurements of specific experiences during relaxation and meditation sessions are presented in Table 1.

During the relaxation and meditation sessions, the subjects most strongly experienced “piece of mind,” “rest,” and “a pleasant feeling” (see Table 1). “Warmth in parts of the body,”

Table 1
Mean Scores for Experiences During
Relaxation and Meditation

Experience	After 5 weeks (n = 205)	After 15 weeks ^a (n = 123)	After 25 weeks ^a (n = 113)
Piece of mind	3.82	3.61	3.78
Rest	4.31	3.98	3.89
The surroundings disappear	3.65	3.24	3.36
Numbness in parts of the body	3.05	2.60	2.89
Feeling of heaviness in the body	3.82	3.34	3.28
Warmth in parts of the body	2.35	2.10	2.24
Disengaged thoughts	3.37	3.07	3.06
Lightness in the body	2.14	1.92	2.28
Tension is released	3.54	3.12	3.41
Tired at awakening	3.31	3.15	3.38
Twitchings in the muscles	1.55	1.46	1.75
Sinking downwards	3.14	2.77	2.96
A pleasant feeling	3.96	3.74	3.82
Difficulty in estimating time	3.80	3.28	3.37

Note. A 5-point scale ranging from *not at all* (1) to *much* (5) was used.
^a Conscript subjects only.

“lightness in the body,” and “twitchings in the muscles” were experienced the least. Table 1 also notes that the strongest experiences occurred at 5 weeks. On three items, the change between the first and second assessments were significant (*t* tests of the significance of the difference between means for correlated samples, two-tailed probability). These items were “numbness in parts of the body,” *t*(117) = 2.51, *p* < .05; “feeling of heaviness in the body,” *t*(117) = 2.99, *p* < .01; and “difficulty in estimating time,” *t*(117) = 3.46, *p* < .01. Significant changes between the first and third assessments were noted on two items: “rest,” *t*(109) = 2.26, *p* < .01, and “feeling of heaviness in the body,” *t*(109) = 3.41, *p* < .01. No significant differences of means were found between the second and third measurements. After 5 weeks the subjects had just completed the basic relaxation training program (Phase 3) and were averaging three relaxation sessions per week. By the time the second and third measurements were taken, the subjects were practicing relaxation or meditation less than 1.5 times per week (Phase 4). The reduced figures could be attributed to this.

At the time of measurement, the subjects were asked to rate and compare their “ability to relax now with prior to training.” A 5-point scale ranging from *much worse* (1) to *much better* (5) was used. The following mean scores were obtained: First measurement = 4.0, second measurement = 3.95, and third measurement = 3.94. The differences between these means were not significant. These figures indicate that the subjects consistently rated their ability to relax as “somewhat better” than it was prior to relaxation training. Interview data also indicated that most subjects (85%) found it difficult to relax or meditate without the help of a tape-recorded program. Because one of the goals of the study was to help the participants become

competent in the independent use of a relaxation or meditation technique, this result was a disappointment.

A total of 95% of the subjects described the content of the relaxation and meditation programs as “good” or “very good.” However, the subjects criticized the program on two points—the low frequency of training from the fifth week onward and the content of the imagery programs. Several subjects claimed that these programs were too structured and did not leave enough time for independent imagery.

According to interview data, the platoon officers experienced their indoctrination training program (Phase 1) as “interesting” and “easy to understand.” Introducing the program to the conscripts and cadets (Phase 2) was considered “not too difficult but requiring time for preparation since it was a new experience.” The training (Phases 3 to 6) was judged by the officers as “easy to lead.” Group sizes varied according to local facilities, but the best results were generally achieved in groups of 15 subjects or less. Like the conscript and cadet subjects, the platoon officers rated the relaxation and meditation programs as “good” or “very good” and the imagery rehearsal training as “too structured.” According to the officers, the major difficulty with the mental training program was finding the time and opportunity to carry it out during field exercises. Field exercises were scheduled about every second week from the fifth week on. This affected the training intensity. Approximately 50% of the platoon officers noted an improved relationship with the conscripts as a possible spin-off effect of the mental training program.

Performance on the Main Tasks

Results on the selected examinations of main task performance are presented in Table 2.

From the results listed in Table 2 it can be concluded that the anti-aircraft artillery conscripts in the experimental groups performed better than did the control subjects on all four examinations. The differences were statistically significant on two of the tests. Table 2 also shows that the conscripts in the experimental groups from the signal troops performed better than their control subjects on slow as well as on fast rates of Morse-code signal receiving. The figures shown in Table 2 for the signal troop conscripts are made up of averages based on 16 independent task examinations. The experimental groups performed better than their control subjects on all tests. The differences were statistically significant on 12 of the examinations. No particular tendencies can be shown between the two groups on standard deviations.

One of the experimental platoons from the signal troops (*n* = 18) did not have a control platoon and its performance on the main task (receiving of Morse code signals) was rated by their radio teacher. This officer rated the platoon as “usual” during ordinary lessons and as “better than usual” during stressful examinations.

The two experimental platoons consisting of cadets only (*n* = 80) did not have a control. Their performance data on the main task (leadership capacity) consisted of self-ratings. On an open-ended questionnaire, 63% of these subjects reported that they had not noticed any differences as to how they had handled stressful situations since beginning mental training. A total of

37% of the subjects reported that they handled these situations better since the start of the training. Many (18 out of 30) in this latter group spontaneously reported "improved concentration ability" when asked "In what way?" Of the subjects who answered "no effect," one third wrote that they considered themselves calm, stable persons, and did not need the training. One third of these subjects wrote that they had not experienced a stressful situation since the start of the training. The remaining subjects did not write anything on the follow-up question, "Why?" None of the cadets reported performing worse in stressful situations since the start of the mental training.

In sum then, on the main tasks the mentally trained subjects performed better than did their control counterparts. Note also that the subjects in the experimental platoons had a 3–5% reduction in actual training time on the main tasks.

Performance on Mental Tests

As reported in detail elsewhere (Larsson, Starrin, Olofsson, & Bäck, 1986), the subjects in the experimental group performed significantly better on mental tests as a long-term effect of relaxation training. The improvement centered on tests that were designed to measure right-hemisphere functions (which was predicted), as well as tests that were designed to measure left-hemisphere functions (which was unanticipated).

Three spatial tests were used to assess right-hemisphere functions. They were (a) a localization test, in which an *X* was marked in a large black frame projected on an overhead picture for 3 s. The subject had to mark the location of the *X* in a similar frame on an answer sheet; (b) a gestalt completion test. A silhouette picture of a rider on a horse (from the Street Gestalt Completion Test; Street, 1931), in which random parts of the picture had been erased, was presented on an overhead picture for 15 s. The subject was to imagine and write down how the completed picture of the object should look; and (c) a conventional spatial ability test consisting of 40 items. Each item consisted of a figure cut into two or more parts. The subject had to determine how the pieces fit together in the complete figure. He then chose the drawing that correctly showed this arrangement.

The left-hemisphere functions were assessed with the following tests: (a) serial numbers, in which seven series of three to nine digits were played on a prerecorded tape at the rate of one digit per second. At the end of each series, the subject was to write the sequence of numbers just as they were played; (b) verbal ability, in which a conventional synonym test including 40 items was used; and (c) verbal fluency, in which the subject was requested to write as many words (Swedish) as possible in a 2-min period that started with the letter *S* and ended with the letter *A*.

As predicted, the experimental group performed significantly better than the control subjects on a test of right-hemisphere functions immediately following relaxation. No differences between the groups were found on a test of left-hemisphere functions immediately following a relaxation session (a performance decrement for the experimental group was anticipated). The study was duplicated with some minor refinements (the first study involved conscript subjects; the second study used cadet subjects). The results on both studies were almost identical and it was concluded that relaxation training (and meditation)

affected performance positively on tests intended to measure basic psychological abilities. Furthermore, the mentally trained group showed a smaller within-group variation on most of the tests.

Results of Mental Training on Well-Being

No significant results were found among the conscript subjects on the questionnaire designed to measure the various aspects of well-being.

Discussion

Mental training among Swedish conscripts and cadets was found to lead to improvements on actual task performance. Subjective reports also indicate that the mentally trained subjects handled stressful situations better than did their control counterparts; increased concentration was mentioned frequently. Performance improvements were also noted on mental tests. Contrary to expectations no effects were found on well-being.

This study differed from most previous research in the following respects: (a) a larger sample size was used, (b) the relaxation training and meditation lasted for a longer period of time, (c) the subjects were above-average physically and mentally, (d) reliable and valid criteria of actual performance were used, and (e) the explicit goal to routinize mental training according to a simple pedagogical model was new.

Most of the goals supporting the routinization of mental training were reached. From the organization's point of view the training was inexpensive, time efficient (about 1% of total training time), and it interfered minimally with ordinary activities. The structure of the training program allowed the individual participant to find the most suitable technique within about 5 weeks. The second goal on the individual level, to learn the most suitable technique thoroughly and use it independently, was not reached. Only a few subjects were able to relax or meditate without the help of a tape-recorded program.

Virtually all of the subjects experienced the relaxation and meditation program positively. In any training program the reaction of the pupil, favorable or unfavorable, is of great importance; it is likely to determine the extent to which he makes use of what he has learned. In the light of this assumption, failure to learn to use a relaxation or meditation technique spontaneously can be viewed as remarkable. However, I tend to think that in a military setting it is difficult to convince conscripts to discipline themselves enough to learn to relax independently. The fact that virtually all program teachers (i.e., platoon officers) experienced the program positively, despite being ordered to carry it out, indicates that teachers of this kind of training do not require continuous moral support from enthusiastic agents. The tendencies of a positive spin-off effect from the training program—improved relationships between officers and conscripts—may have contributed to the positive attitude of the officers.

The most striking result of the study is the increase in performance in light of the modest amount of time and resources the mental training program demanded. It may be suggested that the increase in performance can be attributed to the novelty of

Table 2
Mean Scores on the Performance Examinations

Platoon and military task examination	Group						<i>t</i> ^a
	Experimental			Control			
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>	
Anti-aircraft artillery							
1st platoon							
Easy targets	32	66.55	21.82	32	99.56	25.84	4.40**
Challenge trophy	32	167.33	10.54	32	177.33	7.65	0.77
2nd platoon							
Easy targets	32	58.29	16.41	32	74.70	19.08	2.82**
Battalion commander's test	32	6.13	1.83	32	5.63	1.94	1.01
Signal troops							
1st platoon							
Slow rates	32	87.09	18.51	32	94.80	20.08	1.85*
Fast rates	32	125.75	19.14	32	142.00	17.18	2.90**
2nd platoon							
Slow rates	20	79.12	17.49	20	85.70	16.39	2.27*
Fast rates	20	108.96	12.52	20	121.28	21.65	1.86*

Note. On all examinations but one, the battalion commander's test, scores reflect either number of needed trials, number of penalty points, or amount of needed time; hence, a low score reflects good performance and vice versa. On the battalion commander's test, the scores indicate number of hits, and hence, high scores reflect good performance and vice versa.

^a *t* test of difference between independent group means.

* *p* < .05. ** *p* < .01.

the program and the attention it attracted, independent of its content, and henceforth was a “Hawthorne” effect. It can also be argued that the subjects expected to perform better after having received the training; thus, their motivation increased. It is also possible that the mental training was perceived as a break in traditional military training. The ideal solution would have been a placebo group. In this case, no acceptable placebo was found. Some suggestions were unacceptable because they might yield results similar to relaxation and meditation (e.g., jogging). Other suggestions were unacceptable to the military authorities because they deviated considerably from ordinary military training. This is an understandable objection and one that most researchers in applied contexts have to face. In the absence of a placebo group, I made attempts to detect signs of a Hawthorne effect in the interviews, but no signs were found. One must remember that the only change in military training was adding the mental training program; all other conditions remained the same. Furthermore, even though all of the subjects knew they were taking part in an investigation, the change was not noteworthy because mental training was just another part of their military training. It is my belief that any effect produced by the investigation itself would quickly wear off. In this case the performance examinations were evenly distributed over the entire training period.

A problem in the present study was the low frequency of training; only about 75% of the planned mental training took place. Most of the conscripts, cadets, and officers were positive about the training and were irritated with the reduced training frequency. The major reason for the reduction was that it was difficult to carry out the training during the scheduled weeks of field exercises. A practical suggestion for remedying this prob-

lem that will be tried in a coming study, is placing a small tape recorder in each tent for individual use during field exercises.

The absence of results on well-being variables cannot be attributed to the reduced training frequency because previous studies (e.g., Setterlind, 1983) showed positive results with fewer training sessions. One factor may be the military's selection requirements for the positions of the subjects in the study (i.e., the subjects were physically and emotionally stable at the onset of the study). Also, the subjects were young, about 20 years old. Many stress-related psychosomatic disorders have a higher frequency later in life. Thus, a suggestion for future research is to study the effects of the training program on a sample with mixed background characteristics (e.g., age, sex, level of education).

The pedagogical model required the presence of or easy access to a supervisor with psychotherapeutic training (the author in this case). Note that no episode occurred that demanded this kind of consultation; this is a difficult ethical question. In light of this study, Setterlind's (1983) study, and personal experience with thousands of athletes who have practiced mental training regularly without any negative psychic symptoms, I think it can be safely stated that this training program is virtually free of risk. Hence, the requirement for a psychotherapeutically trained supervisor can probably be deleted. These conclusions are obviously limited to mental training programs involving nonclinical individuals. Among psychiatric patients, relaxation training and meditation must be conducted with great care (see, e.g., Larsson et al., 1985).

The practical value of the performance improvements is difficult to assess. However, given that the criteria for performance are valid, results suggest that training time for the main

task can be reduced by about 3 to 5%. Subjective reports indicate that the mentally trained subjects performed better in stressful situations. Whether these results persist in extremely stressful conditions, like actual combat, remains (hopefully not) to be seen. Further research is needed to evaluate how enduring the results are if one quits practicing and to assess the practical value of the tendencies of smaller within-group variations among the mentally trained.

In conclusion, the suggestion has been put forth to pay particular attention to the imagery segment of the training program. In this study, highly structured tape-recorded programs were used. It was assumed that the conscripts' minds would wander to other subjects if they were simply instructed to imagine freely on their own. However, the highly structured program interfered with the imagery process. Thus, in addition to the original requirements for this type of training, a free program based on metaphorical (right-hemisphere learning) instructions rather than specific, sequentially structured (left-hemisphere learning) instructions, plus a certain level of motivation, are also needed (Cook, 1984). Given that these requirements are fulfilled, the value of imagery rehearsal training, alone or combined with relaxation and meditation, should be systematically assessed in applied contexts. All educational training systems can be looked on as simulation. There are different degrees of simulation or deviances from a thought reality. Imagery or mental simulation may prove to be a valuable complement to, and perhaps reduce the cost of, conventional forms of simulation.

Implementation of the results of this study is currently being carried out by decision makers in the Swedish Armed Forces. Mental training will become part of the educational program for officers; this training will qualify them as future trainers. In 2 or 3 years the final aspect of the program will be realized—several thousand conscripts will learn mental training annually as part of their military basic training.

References

- Anand, B. K., Chhina, E. S., & Singh, B. (1961). Some aspects of electroencephalographic studies in yogis. *Journal of Electroencephalographical Clinical Neurophysiology*, 13, 452–456.
- Appelle, S., & Oswald, L. E. (1974). Simple reaction time as a function of alertness and prior mental activity. *Perceptual and Motor Skills*, 38, 1263–1268.
- Arnold, J. B. (1970). The relationship between hypnosis and the learning of two selected motor skills (Doctoral dissertation, Temple University, 1969). *Dissertation Abstracts International*, 31, 1053A.
- Bakan, P. (1969). Hypnotizability, laterality of eye-movements and functional brain asymmetry. *Perceptual and Motor Skills*, 28, 927–932.
- Bakan, P. (1980). Imagery, raw and cooked: A hemispheric recipe. In J. Shorr (Ed.), *Imagery: Its many dimensions and applications* (pp. 35–54). New York: Plenum Press.
- Barabasz, A. F. (1980). Effects of hypnosis and perceptual deprivation on vigilance in a simulated radar target-detection task. *Perceptual and Motor Skills*, 50, 19–24.
- Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology*, 46, 610–620.
- Benson, H. (1975). *The relaxation response*. New York: Morrow.
- Bhatia, N., & Murrell, K. F. H. (1969). An industrial experiment in organized rest pauses. *Human Factors*, 2, 167–174.
- Blumberger, S. R. (1978). An overview of the method, theory and outcome of the creativity mobilization technique. *Creative Child & Adult Quarterly*, 3, 91–97.
- Brown, B. B. (1970). Recognition of aspects of consciousness through association with EEG alpha activity represented by a light signal. *Journal of Psychophysiology*, 6, 442–452.
- Brown, B. B. (1980). *Supermind: The ultimate energy*. New York: Harper & Row.
- Browne, M. A., & Mahoney, M. J. (1984). Sport psychology. *Annual Review of Psychology*, 35, 605–625.
- Carrington, P., Collings, G. H., Benson, H., Robinson, H., Wood, L. W., Lehrer, P. M., Woolfolk, R. L., & Cole, J. W. (1980). The use of meditation–relaxation techniques for the management of stress in a working population. *Journal of Occupational Medicine*, 22, 221–231.
- Chaney, D. S., & Andreasen, L. (1972). Relaxation and neuromuscular tension control and changes in mental performance under induced tension. *Perceptual and Motor Skills*, 34, 677–678.
- Chaney, D. S., & Andreasen, L. (1973). Relaxation and neuromuscular tension control and changes in motor performance under induced tension. *Perceptual and Motor Skills*, 36, 185–186.
- Clark, L. V. (1960). Effect of mental practice on the development of a certain motor skill. *Research Quarterly*, 31, 560–569.
- Cook, C. M. (1984). *A phenomenological study in the design and application of right hemisphere learning programs in sport*. Unpublished master's thesis, Sonoma State University, Department of Psychology.
- Cox, T. (1978). *Stress*. London: MacMillan.
- Danskin, D. G. (1981). *Biofeedback: An introduction and guide*. Palo Alto, CA: Mayfield.
- Davidson, R. J., & Schwartz, G. E. (1976). The psychobiology of relaxation and related states: A multi-process theory. In D. Mostofsky (Ed.), *Behavior control and modification of physiological activity* (pp. 399–442). Englewood Cliffs, NJ: Prentice Hall.
- Decaria, M. D. (1977). The effect of cognitive rehearsal training on performance and on self-report of anxiety in novice and intermediate female gymnasts. *Dissertation Abstracts International*, 38. (University Microfilms No. 77-14, 292)
- DeMers, G. E. (1980). Effects of post-hypnotic suggestion on the performance of a fine motor skill under stress. (Doctoral dissertation, University of Utah, 1979). *Dissertation Abstracts International*, 40, 4955A.
- DeWitt, D. J. (1980). Cognitive and biofeedback training for stress reduction with university athletes. *Journal of Sport Psychology*, 2, 288–294.
- Galin, D. (1974). Implications for psychiatry of left and right cerebral specialization. *Archives of General Psychiatry*, 31, 572–583.
- Garver, R. B. (1977). The enhancement of human performance with hypnosis through neuromotor facilitation and control of arousal level. *American Journal of Clinical Hypnosis*, 19, 177–181.
- Girdano, D. A., & Everley, G. S. (1979). *Controlling stress and tension: A holistic approach*. Englewood Cliffs, NJ: Prentice Hall.
- Gordin, R. D. (1981). Effects of hypnosis, relaxation training, or music on state anxiety and stress in female athletes (Doctoral dissertation, University of Utah, 1981). *Dissertation Abstracts International*, 42, 598A–599A.
- Hickman, J. L., Murphy, M., & Spino, M. (1977). Psychophysical transformations through meditation and sport. *Simulation & Games*, 8, 49–60.
- Kamiya, J. (1969). Operant control of the EEG alpha rhythm and some of its reported effects on consciousness. In C. T. Tart (Ed.), *Altered states of consciousness* (pp. 507–518). New York: Wiley.
- Klisch, K. (1981). An investigation of a relaxation imagery technique on the performance of a motor skill (Doctoral dissertation, University of Maryland, 1980). *Dissertation Abstracts International*, 41, 3478A.

- Korn, E. R., & Johnson, K. (1983). *Visualization: The uses of imagery in the health professions*. Homewood, IL: Dow Jones-Irwin.
- Larsson, G. (1984). *Manuskript för fördjupningsträning av avslappning* [Manuscripts for deep training of relaxation]. (Report No. PM 55:66). The Swedish National Defence Research Institute, Division of the Behavioural Sciences, Karlstad, Sweden.
- Larsson, G., & Bäck, P. (1985). *Försök med avslappnings—och föreställningsträning vid utbildningen av värnpliktiga skyttar på robot 70 vid Lv 4 1982/83* [Relaxation and imagery rehearsal training with conscript Rbs 70 shooters 1982/83]. (Report No. PM 55:83). The Swedish National Defence Research Institute, Division of the Behavioural Sciences, Karlstad, Sweden.
- Larsson, G., & Setterlind, S. (1983). *Mental träning inom idrotten* [Mental training in sports]. Stockholm, Sweden: Friskvårdscentrums Förlag.
- Larsson, G., Setterlind, S., & Steffner-Starrin, L. (1985). *Avslappningsträning inom hälso-och sjukvård* [Relaxation training in health care]. Stockholm, Sweden: Friskvårdscentrums Förlag.
- Larsson, G., Starrin, B., Olofsson, K., & Bäck, P. (1986). *Effects of relaxation training on verbal ability, sequential thinking, and spatial ability* (Report No. C 50029-H3). The Swedish National Defence Research Institute, Division of the Behavioural Sciences, Karlstad, Sweden.
- Lazarus, R. S. (1966). *Psychological stress and the coping process*. New York: McGraw Hill.
- Mahoney, M. J., & Avenier, M. (1977). Psychology of the elite athlete: An exploratory study. *Journal of Cognitive Therapy and Research*, 1, 135–141.
- Nelson, J. (1980). Investigation of effects of hypnosis, relaxation, and mental rehearsal on performance scores of golfers and runners (Doctoral dissertation, The Louisiana State University and Agricultural and Mechanical College, 1980). *Dissertation Abstracts International*, 41, 1484B.
- Neufeld, W. (1951). Relaxation methods in U.S. Navy Air Schools. *American Journal of Psychiatry*, 8, 132–137.
- Nideffer, R. M., & Deckner, C. W. (1970). A case study of improved athletic performance following use of relaxation procedures. *Perceptual and Motor Skills*, 30, 821–822.
- Østbye-Sundsvold, M., & Vaglum, P. (1984). Muscular pains and psychopathology. In T. H. Hoskins (Ed.), *International perspectives in physical therapy: Volume on pain* (pp. 1–32). London: Churchill Livingstone.
- Pardine, P., Dytell, R., & Napoli, A. (1981). Transfer benefits of biofeedback: A research note. *Perceptual and Motor Skills*, 52, 373–374.
- Pelletier, K. R. (1974). Influence of transcendental meditation upon autokinetic perception. *Perceptual and Motor Skills*, 39, 1031–1034.
- Peters, R. K., Benson, H., & Porter, D. (1977). Daily relaxation response breaks in a working population: I. Effects on self-reported measures of health, performance and well-being. *American Journal of Public Health*, 67, 946–953.
- Richardson, A. (1983). Imagery: Definition and types. In A. Sheikh (Ed.), *Imagery: Current theory, research, and applications* (pp. 3–42). New York: Wiley.
- Rubenzon, R. (1979). The role of the right hemisphere in learning & creativity implications for enhancing problem solving ability. *The Gifted Child Quarterly*, 23, 78–100.
- Sage, G. H. (1971). *Introduction to motor behavior: A neurological approach*. Reading, MA: Addison-Wesley.
- Scott, M. D., & Pellicioni, L. (1982). *Don't choke: How athletes can become winners*. Englewood Cliffs, NJ: Prentice Hall.
- Setterlind, S. (1983). *Avslappningsträning i skolan: Forskningsöversikt och empiriska studier* [Relaxation training in school: Overview of research and empirical studies]. Göteborg, Sweden: Acta Universitatis Gothoburgensis.
- Setterlind, S., & Larsson, G. (1983). *Må bättre genom avslappning* [Feel better with relaxation]. Stockholm, Sweden: Friskvårdscentrums Förlag.
- Shaw, W. A. (1940). The relation of muscular action potentials to imaginal weight lifting. *Archives of Psychology*, 237, 50.
- Springer, S. P., & Deutsch, G. (1982). *Left brain, right brain*. San Francisco, CA: Freeman.
- Street, R. F. (1931). *A gestalt completion test: A study of a cross section of intellect* (Teachers' College Contributions to Education No. 481). New York: Columbia University, Teachers College.
- Suinn, R. M. (1980). Psychology and sports performance: Principles and applications. In R. M. Suinn (Ed.), *Psychology in sports: Methods and applications* (pp. 3–23). Minneapolis, MN: Burgess.
- Uneståhl, L-E. (1979). *Självkontroll genom mental träning: Tillämpningar idrott* [Mental training for self-control: Sports applications]. Örebro, Sweden: Veje Förlag.
- Weinberg, R. S. (1982). The relationship between mental preparation strategies and motor performance: A review and critique. *Quest*, 33, 195–213.
- Williams, L. R. T., & Herbert, P. G. (1976). Transcendental meditation and fine perceptual-motor skill. *Perceptual and Motor Skills*, 43, 303–309.
- Williams, L. R. T., & Vickerman, B. L. (1976). Effects of transcendental meditation on fine motor skill. *Perceptual and Motor Skills*, 43, 607–613.
- Wolpe, J. (1958). *Psychotherapy by reciprocal inhibition*. Stanford, CA: Stanford University Press.

Received September 27, 1985

Revision received July 10, 1986 ■

Effects of Categorization, Attribution, and Encoding Processes on Leadership Perceptions

Steven F. Cronshaw

University of Waterloo, Waterloo, Ontario, Canada

Robert G. Lord

University of Akron

In this study we compared two cognitive processes that are often thought to precede leadership perceptions: causal attributions and categorization. This was done by experimentally manipulating factors relevant to attributions (consensus information) and categorization (stimulus prototypicality). Dependent measures were undergraduate subjects' perceptions of the leadership exhibited by stimulus people, shown on a 12-min videotape of a management group. The interaction of the leader prototypicality and consensus information factors on leadership perceptions was opposite to that predicted by attribution theory. The experimental evidence suggested that the interaction effect was based on subjects' categorization of stimuli in terms of leadership. A methodology developed to measure encoding of on-going leader behavior allowed tests of the social-information-processing sequence involved in forming leadership perceptions. Results support recent propositions of social-information-processing theory and demonstrated the usefulness of the encoding methodology.

The topic of leadership perceptions is of interest to scientists with both theoretical and applied orientations. Theoretically, leadership perceptions reflect the broader processes used to form many types of social perceptions, and they can be analyzed in terms of underlying social or cognitive processes. Practically, leadership perceptions involve key interpersonal processes in organizations that impact on the formation of status or influence structures and the development of superior-subordinate relations (Seers & Graen, 1984).

We investigated this important topic by comparing two cognitive processes that are often thought to be immediate antecedents to leadership perceptions, namely, categorization and attributional processes. Social categorization (Cantor & Mischel, 1979) and attributional (Kelley, 1973) theories are the two major theoretical approaches to understanding person judgments that are current in the social cognitive literature. Yet, they remain largely independent research domains, although some preliminary work has established interesting interrelations between these constructs in the leadership area (Phillips & Lord, 1981). We extend this work by comparing the impact of categorization and attributional processes on leadership perceptions

in a laboratory study. Thus, the study attempts to further our understanding of leadership perceptions and also to contrast these two alternative theories of social perceptions.

Although we focused specifically on leadership perceptions, categorization and attributional principles are also incorporated into models used for understanding performance appraisal (Feldman, 1981). In these models, both person impressions and causal attributions influence how raters acquire information about ratees, but attributional processing may precede, and largely determine, impression formation (DeNisi, Cafferty, & Meglino, 1984; Ilgen & Feldman, 1983). However, causal attributions to internal or external causes have also occurred after people categorize others as "ingroup" and "outgroup" members (Wilder, 1981), indicating that categorization can precede attributions. Linkages between categorization and attributional processes are further explored in articles discussing stereotyping (Hamilton, 1979), attribution theory (Kelley, 1972), and social cognition (Hastie, 1981). In addition, other work applies Kelley's (1973) attributional principles to explain superiors' reactions to subordinate performance (Green & Mitchell, 1979). However, the alternative explanations of categorization and attribution processes in determining person perceptions have not been experimentally compared in a single study within either the leadership or related organizational or social psychology literature.

Categorization Theory

Lord and his colleagues (Lord, Foti, & De Vader, 1984; Lord, Foti, & Phillips, 1982) described how categorization can operate to determine leadership perceptions. Certain salient features or behaviors of the leader initiate a limited search for the category prototype that matches those features or behaviors, where the *prototype* is a set of characteristics possessed by most category members. For example, a prototypical leader is decisive, intelligent, and industrious (Lord et al., 1984). If a match to a leadership prototype is made, a leader label is applied to

This study is based on the first author's doctoral dissertation under supervision of the second author.

The first author thanks the following committee members for their advice and guidance: Gerald V. Barrett, Ralph Alexander, and Jonathan Smith. Mark Porter of the University of Akron's electronics shop provided valuable assistance in building and installing the response recorder used in this study. Further thanks go to the following people for their assistance during development of the stimulus videotapes and for providing other essential technical resources: Nick Barry, Jon Loufman, George Graham, Eric Kreider, Andy Murray, Al Coursol, and Ron Gaug. We also thank Mark Zanna for his comments on an earlier draft of this article.

Correspondence concerning this article should be addressed to Steven F. Cronshaw, who is now at the Department of Psychology, University of Guelph, Guelph, Ontario, Canada, N1G 2W1.

the stimulus person and is then stored in long-term memory. The perceiver can then use this leader label to access the corresponding leader prototype when asked to make judgments concerning the stimulus person. Thus, categorization is a simplifying heuristic that reduces encoding and memory demands. Several empirical studies support the contention that observers do use a categorization process to form leadership perceptions (Cronshaw & Lord, 1982; Foti, Fraser, & Lord, 1982; Fraser & Lord, 1984; Lord et al., 1984; Phillips, 1984; Phillips & Lord, 1982).

Attribution Theory

Attributional processes may demand more controlled or effortful processing by observers than does categorization (Lord & Smith, 1983). For example, Pfeffer (1977) pointed out that leadership is attributed by observers to salient persons in the organization, thereby emphasizing personal rather than environmental control of organizational performance. In his theory, attributional reasoning precedes and largely determines leadership perceptions. Calder (1977) presented a more detailed four-stage attributional model of leadership. The important initial linkages in Calder's model are in fact examined in this study. Calder began by assuming that individuals observe behavior by others as well as observing the effects of these behaviors. The observed actions and effects may imply other behaviors that were never observed. In the second stage of the attribution process, actions and effects associated with the focal person (i.e., the leader) are compared with behavior of other actors in the group. If a considerable amount of variability in behavior is found, a leadership inference is made. If little variability is found, inferences about leadership are less likely. When a leadership inference is made, it is matched to a "typicality inference" (the equivalent of a leader prototype). According to Calder, the typicality inference is then subjected to further analysis such as determination of social desirability of the behavior exhibited. However, the group comparison-typicality inference linkage is the important primary stage in social information processing proposed by Calder in which attributional reasoning precedes leadership perceptions.

Calder (1977) drew heavily on the attribution literature for his model. This literature in turn explains in more detail how the group comparison-typicality inference linkage can impact on leadership perceptions. Group comparison is in fact synonymous with consensus as used by attribution researchers. McArthur (1972, 1976) found that covariance information, including consensus, determined whether subjects made attributions to persons or to other causes (e.g., situations). She defined consensus as "whether or not the same response is produced by other people in the presence of the entity" (McArthur, 1972, p. 172), and manipulated consensus through written statements describing "almost everyone" or "hardly anyone" behaving in the same manner as the focal person. Other research has shown that when subjects made greater *personal* attributions to a focal person, the impact of information on subjects' impressions was increased (Crocker, Hannah, & Weber, 1983).

The preceding attributional research therefore yields the same predictions as does Calder's (1977) model, although it also incorporates the concept of personal attribution. When a leader

is operating under low consensus (i.e., when behavior is variable over the group because the leader behavior is inconsistent with others), observers will form stronger leadership impressions. Attribution theory further suggests that this stronger leadership impression results from the greater dispositional attributions made under low consensus.

Three Process Models

The previous discussion presents categorization and attribution models as alternative explanations for the formation of leadership perceptions. This section has two purposes. First, we identify a priori three alternative process models that specify how categorization and attributional reasoning might impact on leadership perceptions. Second, these models are translated into specific predictions at the operational level. These operational predictions are easily derived because we directly manipulate the categorization and attributional constructs. Use of categorization is manipulated by varying the prototypicality of leader behavior viewed by subjects. This use of the prototypicality manipulation is consistent with a series of leadership categorization studies (Fraser & Lord, 1984; Lord et al., 1984). Attributional reasoning is manipulated by varying consensus information, as suggested by Calder's (1977) model.

Model 1, which we call the *categorization model*, assumes that leadership perceptions are based primarily on the leader categorization process. Under this model, attributional reasoning is not used by observers in forming leadership perceptions. Operationally, the model would be supported if only the leader prototypicality manipulation impacts as a main effect on subsequent leadership ratings.

Model 2, which we call the *independent model*, assumes that observers use both categorization and attributional reasoning in parallel fashion when forming leadership perceptions. That is, both prototypical behavior and personal causation have independent effects on leadership perceptions. Operationally, Model 2 would be supported if the leader prototypicality and consensus information manipulations impact as independent main effects on leadership ratings.

Model 3, which we call the *attributional model*, assumes that attributional reasoning is required to interpret stimulus behavior. According to this model, internally caused behaviors would be given much more weight than externally caused behaviors in forming leadership perceptions. Operationally, Model 3 would be supported by an interactive effect of leader prototypicality and consensus information on leadership ratings. The previous consensus model clearly showed that observers should form a more extreme person impression of the stimulus person under low as compared to high consensus. Based on these results, we would predict that the impact of the prototypicality manipulation would be greater under low than high consensus conditions. This hypothesized interaction is diagrammed in the left-hand panel of Figure 1.

Encoding and Leadership Perception

This study will also investigate the effects of subjects' stimulus encoding on leadership perceptions. Although schema labeling and retrieval processes have been investigated in previous

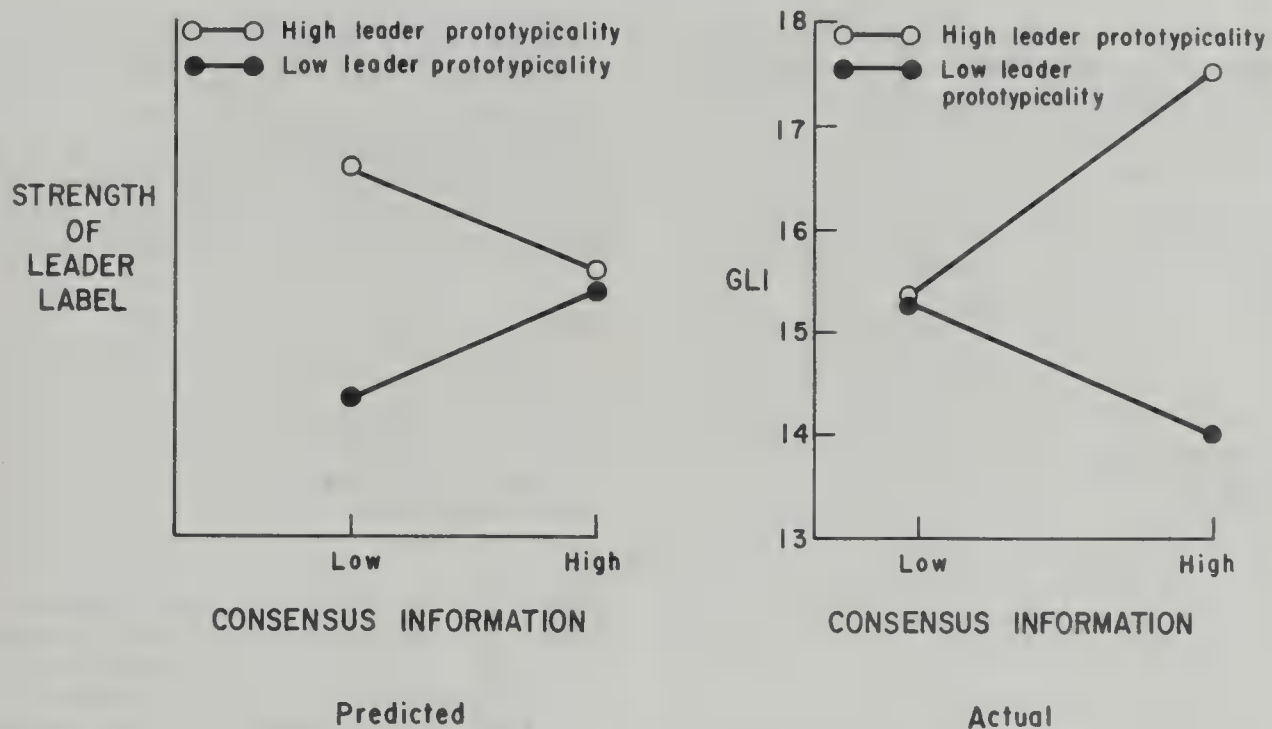


Figure 1. Predicted and actual interaction effect of leader behavior prototypicality and consensus information manipulations of leader labeling.

leadership categorization work (Phillips, 1984; Phillips & Lord, 1982; Rush, Phillips, & Lord, 1981), our study is the first to examine the impact of leader behavior encoding on leadership perceptions. Thus, we extend work that assesses encoding of nonverbal behavior (Newton, 1973; Newton, Engquist, & Bois, 1976, 1977; Newton, Rindner, Miller, & LaCross, 1978) and the effect of self-schemata on encoding of schematic information (Markus & Smith, 1981).

In addition, we use perceivers' ongoing encoding of behavior to directly measure the cognitive constructs we manipulated experimentally. More specifically, we develop process-based measures of consensus encoding and prototype encoding. These measures can then serve as intervening or mediating variables in explaining leadership perceptions. For example, the consensus manipulation can be related to *measured* consensus encoding, which in turn can be related to leadership perceptions. This procedure can corroborate the suggested means by which experimental factors impact on dependent variables.

In summary, this study manipulated leader prototypicality and consensus information in a 2×2 (Stimulus Prototypicality [high or low] \times Consensus [high or low]) factorial design, with encoding and attribution process measures collected to serve as potential mediating variables (covariates) in evaluating impact of categorization and attribution processes on leadership perceptions.

Method

Subjects

Subjects were 104 undergraduates from the University of Akron. An equal number of men and women were distributed across conditions in order to avoid confounding sex of rater with experimental factors. Otherwise, assignment of subjects to conditions was random. Data for

5 subjects were excluded due to equipment failure causing loss of all encoding data.

Stimulus Materials

Four 12-min color videotapes served as the stimulus materials. All four videotapes showed the same target person interacting with two other members of a work group, and were carefully matched for length of running time as well as for total number of words and behaviors engaged in by each actor. The proportions of leader-related or total behaviors and word counts for behaviors of each type were matched between the four stimulus tapes for each of the three group members. For instance, in all conditions there was an average of 23 words for each prototypical or antiprototypical behavior. Finally, the content and logical flow of events was highly similar in all four videotapes. This careful development and balancing of the stimulus tapes was intended to ensure that only the focal constructs of this study (i.e., leader prototypicality and consensus information) were manipulated via the different videotapes.

Professional actors played the parts of the target person and one other group member, and a graduate student played the part of the third person in the group. The work-group members discussed previous financial reports for their manufacturing firm and formulated plans for future plant operations in order to make recommendations to their Board of Directors. In all conditions, performance information was intentionally ambiguous because it affects leadership perceptions and attributions (Phillips & Lord, 1981).

Experimental Manipulations

Leader prototypicality. Leader prototypicality was manipulated by scripting varying numbers of prototypical and antiprototypical leader behaviors into the videotapes. These behaviors were initially identified by combining the prototypicality ratings for 110 leader behaviors that were obtained in four previous studies (Foti & Lord, in press; Fraser & Lord, 1984; Fraser, Lord, & Cronshaw, 1983; Lord et al., 1984). In the high leader-prototypicality condition, the target engaged in a relatively

Table 1
Prototypical and Antiprototypical Behaviors Incorporated into the Stimulus Videotapes

Prototypical	Antiprototypical
1. He (the work-group supervisor) delayed action on decisions (6.34).	1. He (the work-group supervisor) worried over other group members' ideas (2.05).
2. He carefully planned what to do (6.06).	2. He wanted his own way on issues (1.41).
3. He emphasized the group's goals (6.14).	3. He was confused about an issue (1.77).
4. He coordinated the group's activities (6.04).	4. He refused to explain his actions (2.32).
5. He let other group members know what was expected of them (5.94).	5. He let other group members decide what to do (2.03).

Note. Behaviors were repeated in some conditions to yield appropriate frequencies for the prototypicality manipulation. Prototypicality ratings on a 7-point scale are given in parentheses.

large number (20) of prototypical leader behaviors relative to antiprototypical behaviors (5). In the low leader-prototypicality condition, he engaged in a small number (5) of prototypical leader behaviors but exhibited a large number of antiprototypical behaviors (20). The prototypical and antiprototypical behaviors built into the videotapes are presented in Table 1. It is important that an examination of the types of leader behavior reveals that they correspond to a prototype or categorization-based conception of leadership rather than "democratic-autocratic" or "structuring-consideration" dimensions of leadership.

Consensus information. Consensus information was manipulated by varying other group members' prototypical or antiprototypical behavior relative to the target. In the high-consensus condition, both the target and other group members exhibited similar proportions of both prototypical and antiprototypical leader behavior. In the low-consensus condition, the target and other group members had the opposite proportions of both prototypical and antiprototypical behavior. For instance, in one low-consensus cell, the target engaged in 5 prototypical and 20 antiprototypical behaviors, whereas the first group member engaged in 20 prototypical and 5 antiprototypical and the second group member engaged in 12 prototypical and 3 antiprototypical behaviors.

Procedure

Subjects reported to the psychology department for testing in groups of up to 4. They were informed that they would be asked to watch a videotape of a group. As in many real-world situations, the subjects were not told that they would be asked to rate actors' behaviors at a later time.

The subjects were first asked to view a 3-min practice tape and depress the hand-held button every time they saw a "meaningful behavior." After they were familiar with the button-pressing task, the subjects were asked to view one of the four stimulus tapes under similar instructions. Subjects were told that the videotape was taken at a training seminar attended by managers of a large corporation. No attempt was made to direct the subjects' attention to the person whom we refer to as the target. No references to "leadership" or "supervisor" were made at all, requiring the subjects to infer leadership from an initially neutral situation. The present experiment therefore examines leadership emergence where no formal status differences among group members exist initially. Where a priori status or power differentials exist between leader and follower, this study generalizes less well.

Following the stimulus tape, subjects were asked to complete individ-

ual difference measures of cognitive complexity (Bieri et al., 1966) and the tendency to make personal versus environmental causal attributions (Lowe, Medway, & Beers, 1976) in order to impose a 20-min time delay between the stimulus tape and the completion of stimulus ratings, thereby minimizing the effects of veridical recall of leader-related information (a condition again consistent with real-world rating situations). The subjects were then asked to complete the causal attribution and leadership measures. They were then debriefed and dismissed.

Encoding Apparatus and Encoding Measurement Technique

Subjects viewed the tapes on an RCA 25-in. color TV monitor from individual viewing booths that were equipped with audio headphones to eliminate any sound intrusion from other subjects during the same session. Subjects were told to depress a hand-held button when they saw on-going "meaningful" behavior occurring on the videotape. Tone markers, inaudible to the subjects, were placed on a separate videotape sound track to bracket the location of all prototypicality- and consensus-related verbal behaviors.¹ Times of both subject button presses for "meaningful" behaviors and tone markers were collected automatically by a response recorder coupled to a PDP 11/10 minicomputer.

The button-pressing task as a technique to measure encoding of leader prototypicality and consensus information was adapted from Newton et al.'s (1976, 1977) technique for unitizing on-going behavior. In our use of this technique, subject's button presses were thought to indicate the location of encoded information. This approach differs from Newton's, in which unitization is assumed to reflect the form of encoded information (Cohen, 1981). Also, Newton and his colleagues assessed nonverbal action sequences, whereas we measured the encoding of verbal behavior exhibited in a group setting.

The definition of what constituted a "meaningful" behavior was intentionally left ambiguous in order to assess subjects' natural encoding of stimuli without the possible biasing effect of experimenter-supplied labels. Few subjects reported any difficulties in completing the button-pressing task. Upon debriefing, most subjects could, in fact, report the type of strategies they used to identify meaningful behaviors. It is interesting that leadership was an often-mentioned heuristic used by subjects.

Dependent Variables

General leadership impression (GLI). The GLI was the primary measure of leadership perceptions. It was composed of subjects' ratings on 5-point scales indicating (a) the amount of leadership the ratee exhibited, (b) how willing the rater would be to choose the ratee as formal leader, (c) how typical the ratee was of a leader, (d) to what extent the ratee engaged in leader behavior, and (e) the degree to which the ratee fit their image of a leader. These five items were summed to produce a composite GLI measure (coefficient $\alpha = .87$).

Encoded leader behavior. Frequencies of encoded prototypical and antiprototypical leader behaviors were determined separately for the

¹ Latency times required to set tone markers around the relevant stimulus behaviors for the final study were also obtained from a pilot study in which 15 subjects viewed a videotape from a previous study of leadership behavior (Fraser & Lord, 1984) and were asked to stop the tape using a remote control unit every time a "meaningful" behavior occurred. These latencies indicated that setting tone markers 3 s after the onset of the relevant behavior and 2.5 s after completion of that same behavior were most efficient and would "capture" 67% of subjects' button presses that were intended to signal encoding of specific verbal behaviors.

three group members by summing the number of button presses made for each behavior type for each group member by subjects viewing the stimulus videotape.

Encoded consensus information. Indices of consensus encoding were developed by noting where adjacent target- and other-group-member behaviors in the stimulus tape were either similar or dissimilar in terms of leadership prototypicality. Button presses for these shifts in group member interactions were then summed to yield indices of similar (high-consensus) and dissimilar (low-consensus) encoded behaviors. High- and low-consensus indices were computed for each of the two other group members yielding four encoded consensus indices.

Open-ended causal ascription. Subjects were asked to recall the reasons for making their questionnaire ratings for the target and to record these reasons in the order used in making the ratings. Subjects were supplied with a lined blank sheet on which to record their responses. The responses were coded into two categories: (a) attributional responses referring to the ratee's characteristics or behaviors, and (b) attributional responses using consensus information. Attributional responses that were not consistent with either of these two categories were not coded. The interrater reliability coefficients for the open-ended attribution measure were as follows: .92 for attributional responses referring to characteristics or behaviors, .57 for attributional responses referring to consensus information, and .94 for combined behavior and attributional responses.

Closed-ended causal ascription (CA). The CA questionnaire used by Phillips and Lord (1981) measured the extent to which the target was perceived as causal for group performance through a composite rating from ten 7-point Likert scale items. Included in this composite rating were items assessing the extent to which the target's ability, motivation, effort, guidance, and structuring were important causal factors. Coefficient alpha estimates for causal ascription were .86 for this study and .89 in the Phillips and Lord study.

Results

Test of Three Process Models

A correlation matrix for dependent variables is presented in Table 2. Most notable is the high correlation between encoded antiprototypical and consensus behaviors. This result suggests that antiprototypical and consensus behaviors are similarly encoded. According to the schema pointer plus tag model (Graesser, Woll, Kowalski, & Smith, 1980), observers undertake more thorough processing of schema-inconsistent than schema-consistent behavior. Observers may devote some of the increased effortful processing associated with encoding of antiprototypical behaviors to collection of baseline leader behavior data from other group members as well. Summary descriptive statistics and analyses of variance for experimental factors appear in Tables 3 and 4.

With the supervisor GLI composite as the dependent variable, there was a significant main effect of the leader prototypicality manipulation. Consistent with Model 1, the target person received a higher GLI rating in the high-prototypicality condition than in the low-prototypicality condition. The main effect for the consensus factor was nonsignificant. In addition, a significant interaction effect was obtained, but the nature of this interaction was opposite to the predictions from Model 3 (see the right half of Figure 1). The effects of the prototype factor were greatest under high consensus, whereas an attributional model predicts greater effects under low consensus.

No statistically significant effects were obtained for the de-

pendent variable closed-ended CA as shown in the bottom line of Table 4. Thus, contrary to previous attributional work, low consensus did not produce more internal attributions than high consensus, perhaps because our manipulation was audio-video rather than written. A written manipulation may induce controlled processing of attributional information using complex causal schema (e.g., Kelley, 1973), whereas the audio-video manipulation may induce subjects to make simplified or common sense attributions based on dispositional information readily available through the leader prototype (Hansen, 1980). This explanation is supported by the results shown in Table 3, in which the highest and lowest cell means for CA occurred in the same cells as the highest and lowest GLI means. Thus, the experimental manipulations seem to have impacted on CA in a manner similar to the GLI dependent measures, but the effect was too weak to achieve statistical significance.

As a whole, these results provide no indication that dispositional causal ascriptions are prerequisites to leadership perceptions. Although they suggest that prototypical behavioral patterns do affect leadership perceptions, predictions from our initial categorization model (Model 1) do not provide an adequate explanation of empirical findings because they do not predict the interaction with the consensus factor. As a means to better understand these findings, we covaried several encoding and self-report measures relevant to categorization and attributional processes. If these measures adequately tap processes that mediate (intervene) between experimental factors and leadership perceptions, then covariation analysis should eliminate the significant effects of experimental factors on leadership perceptions. This application of analysis of covariance (ANCOVA) to test mediational hypotheses is more fully discussed by Overall and Woodward (1977, pp. 592-593).

The first set of covariates entered are the measures of prototypical and antiprototypical leader behavior encoding for the target (see the first ANCOVA in Table 4). As previously mentioned, leader behavior encoding should be an initial state in categorization on which the leader label is based. This covariate eliminated the main effect of the leader prototypicality factor but not the interaction term. When the encoding measure for total behavior (comprising leader behavior and all other behavior encoding for the target person) is covaried in the second ANCOVA, no corresponding reduction in significance occurs. This finding suggests that leadership perceptions were not merely based on perceived participation rates as found by Stein and Heller (1979) and more recently by Lord and Alliger (1985). Rather the *type* of behavior exhibited is important for determining leadership perceptions.

Two additional covariates, encoded consensus information and the open-ended attributional measure, produced no appreciable drop in significance in either experimental effect (see the third and fourth ANCOVAs in Table 4). These results suggest that neither subjects' encoding nor their reported use of consensus information was responsible for the effects of the experimental manipulations on leadership perceptions.

In short, these ANCOVAs clearly show that total activity and encoded (or reported) consensus behaviors are not mediating variables; that is, the effects of experimental manipulations on leadership perceptions do not operate through these variables. On the other hand, encoded leader behavior is a plausible medi-

Table 2
Correlation Matrix of Dependent Variables

Variable	1	2	3	4	5	6
1. General leadership impression	—	.23**	-.03	.04	.26**	.25**
2. Encoded prototypical leader behavior		—	.15	.22**	.09	.20*
3. Encoded antiprototypical leader behavior			—	.75***	-.06	.15
4. Encoded consensus information ^a				—	.09	.12
5. Open-ended attribution					—	.15
6. Closed-ended attribution						—

^a The single index of encoded consensus information is computed by combining the four encoded consensus indices described in the Method section. Therefore, the summated index represents the total frequency of button presses for both high- and low-consensus behavior in which interaction shifts from the target to either of the other two group members.
* $p < .05$. ** $p < .01$. *** $p < .001$.

ator for the target prototypicality factor, but not for the interaction with the consensus factor.

Relation of Encoding to Leadership Perceptions

In addition to their uses as covariates, encoding measures can be used in regression analyses to predict leadership perceptions. Such a demonstration adds to the categorization literature that has not previously used encoding measures. In Table 5, leadership perceptions are regressed on the prototypical and antiprototypical behavior encoding measures for the target person. In the last row of Table 5, the regression of the GLI composite against the prototypical- and antiprototypical-leadership encoding measures yielded a significant positive beta weight for the prototypical behavior and a negative but nonsignificant weight for antiprototypical behavior.

One explanation for the small proportion of variance explained may be that the relation of encoding to leadership perceptions changed over conditions. That is, when subjects view a large number of prototypical behaviors in the high-prototypicality condition, they should utilize a leader schema to a great extent. The increased amount of prototype-driven processing

should be reflected by increased encoding of leader-relevant behavior. Conversely, encoding of leader behavior should be less apparent in the low-prototypicality condition in which the predominance of antiprototypical behaviors inhibited the emergence of a leadership schema during the videotape presentation. To explore this possibility, separate analyses were performed within the high- and low-prototypicality conditions. The results of the regression analyses presented in the first two rows of Table 5 support our interpretation of the encoding construct. This relation of behavior encoding to leadership perceptions was much greater in the high leader-prototypicality condition.

In summary, the findings in this section supplement those of the previous section. Not only did leader behavior encoding impact directly on leadership perceptions, but the direction of the relations for the respective encoding indices was consistent with categorization theory. These findings therefore offer additional support for the leadership categorization model of leadership perceptions.

Discussion

This study contrasted three alternative models detailing the impact of categorization and attribution on subsequent leadership perceptions. Although the main effect of leader prototypicality was consistent with the first categorization-based model, an interaction effect was also found that was not predicted by the categorization model and was opposite to that expected under the attribution-based Model 3. Measures of categorization and attribution processes were covaried from these effects to further clarify which process was responsible for the obtained results. These ANCOVA findings implicated categorization, rather than attribution, as the determinant of subsequent leadership perceptions, but they did not fully explain our findings. Correlational analyses also showed that encoding of prototypical and antiprototypical stimulus information predicted leadership perceptions.

One possible explanation of the Prototype \times Consensus interaction is that the high-consensus condition served to “prime” the leadership construct in the high-prototypicality condition, but primed the contrasting nonleader construct in the low-prototypicality condition. In other words, the type of behavior of the two other group members either made leadership more or less available to perceivers, thereby increasing or

Table 3
Descriptive Statistics for General Leadership Impression (GLI) and Causal Ascription (CA)

Dependent measure	High target prototypicality		Low target prototypicality	
	Low consensus	High consensus	Low consensus	High consensus
GLI ^a				
<i>M</i>	15.38	17.56	15.31	14.08
<i>SD</i>	4.18	4.23	4.19	4.77
CA				
<i>M</i>	54.58	56.72	56.12	52.23
<i>SD</i>	11.47	9.02	9.63	11.20

Note. $N = 103$.
^a Phillips and Lord (1981) used a single-item GLI assessing perceptions of leadership exhibited. Use of the corresponding single-item GLI in the preceding analyses yielded a pattern of results analogous to the GLI composite. All further analyses therefore use the GLI composite.

Table 4
Analyses of Variance and Covariance for Leadership Perceptions

Dependent measure	Effect					
	Target prototypicality (A)		Consensus information (B)		AXB	
	F	ω ²	F	ω ²	F	ω ²
General leadership impression (covarying)	4.20*	.03	0.26	.00	3.96*	.03
Leader behavior encoding	1.23	.00	0.15	.00	3.83*	.03
Total behavior encoding	5.05*	.04	0.18	.00	4.26*	.03
Consensus behavior encoding	5.44*	.04	0.44	.00	5.43*	.04
Reported use of consensus information	3.89*	.02	0.50	.00	6.75**	.05
Causal ascription to the leader	0.50	.00	0.20	.00	2.16	.01

* *p* < .05. ** *p* < .01.

decreasing their tendency to categorize the target person in terms of leadership. This priming explanation is plausible because priming manipulations have been shown to affect social judgments in several other studies (Higgins & King, 1981; Higgins, Rholes, & Jones, 1977; Srull & Wyer, 1979). Moreover, priming can operate through automatic processes (Bargh & Pietromonaco, 1982), so priming effects could occur for stimulus behaviors not explicitly recognized as being important. Such behaviors would not be picked up by our encoding measure. Also, in a highly related study, Hauenstein and Lord (1986, October) showed that priming leadership through an unrelated task affected leadership perceptions. They used a Bayesian model to explain social judgments, equating priming with the effects of base rate information, and found support for a categorization-based analogy for Bayesian decision making. Because the interaction effect found here was unexpected, we did not incorporate any measures that could empirically test this priming explanation.

Another possible explanation for this interaction is that subjects inferred group performance levels from the combined prototypicality of all subjects' behavior (prototypical, high consensus > low consensus conditions > antiprototypical, high consensus). If so, performance levels could then be used to infer the amount of leadership exhibited, inasmuch as past research has shown that performance feedback affects leadership perceptions (see Lord, 1985, for a review of such results). We did not anticipate this interaction; hence, we included no estimates of

group performance that could be used to evaluate this possibility. Future research, however, should compare the priming and performance feedback explanations of this interaction.

A third possible explanation for the Prototype × Consensus interaction is that subjects based their leadership perceptions on the *total* amount of prototypical and antiprototypical leader behavior exhibited by the three group members rather than just the behavior of the target person they rated. That is, subjects were sensitive to the type of behavior that occurred but not to which stimulus person exhibited the behavior. The high-target-prototypicality–high-consensus condition contained the largest number of total prototypical behaviors, the low-target-prototypicality–high-consensus condition had the fewest, and the other two conditions contained a moderate number of total prototypical behaviors; whereas, this pattern was reversed for antiprototypical behaviors. This pattern of total prototypical leader behavior was similar to the ordering of mean leadership ratings across the experimental conditions. The third explanation was therefore tested by covarying the *total encoded* prototypical and antiprototypical leader behavior for all three group members from the interaction effects reported for leadership perceptions in Table 4. The interaction effect remained significant (*F* = 6.71, *p* < .01). This finding suggested that total encoded leadership behavior for all group members was *not* responsible for the interaction effect, making the third explanation unlikely.

Although it conflicts with the attributionally based models of leadership perceptions such as the one offered by Calder (1977),

Table 5
Effects of Leader Behavior Encoding Frequency on Leadership Perceptions

Condition	<i>n</i>	Prototypical			Antiprototypical			<i>R</i> ²	<i>F</i>
		<i>r</i>	β	<i>t</i> value	<i>r</i>	β	<i>t</i> value		
Leader prototypicality									
High	51	.15	.35	2.28**	-.24*	-.41	-2.67**	.15	4.17*
Low	52	.24*	.28	1.37	.16	-.05	-0.26	.06	1.56
All conditions	103	.23**	.24	2.42**	-.03	-.07	-0.69	.06	2.96

Note. *R*² and *F* values are for the complete regression equation.
* *p* < .05. ** *p* < .01.

our lack of support for the attributional models agrees with several recent studies of social cognitions. Most relevant is the work of Smith and Miller (1983) in which they used reaction time measures to assess which of several types of social judgments were more fundamental. Reasoning that antecedent processes *must* have faster reaction times than subsequent processes, they found much faster reaction times for sex and trait judgments than for causal attribution judgments. Such results are inconsistent with theories in which attributional judgments must first precede trait ascriptions. But they are wholly consistent with models in which causal judgments are derived from already-formed trait ascriptions such as leadership perceptions.

Our findings and explanation should not be interpreted to mean that attributions are never important in forming leadership perceptions. They simply imply that attributional information may not be picked up from ongoing behavioral episodes and that attributional reasoning is not required to integrate this information into trait ascriptions. There are other instances in which attributional reasoning is clearly important in forming leadership perceptions. For example, Phillips and Lord (1981) found that performance feedback had a much larger impact on leadership ascriptions when coupled with information implying that the prospective leader was causally important, than did similar performance information when coupled with information implicating situational causality. Such theorizing suggests two ways that leadership perceptions can be formed. *Recognition based processes*, which are largely automatic and emphasize categorization processes, can be used to select and interpret ongoing social information as was done in the present study. *Inferential based processes*, which are more controlled and emphasize attributional reasoning, may be used to reflect on past events and integrate specific, highly salient information such as performance information (see Lord, 1985, for a thorough discussion of these alternative bases of leadership perceptions).

One final alternative to our categorization interpretation is that subjects formed an overall impression of the target in terms of favorability and derived their leadership ratings from this impression. Fortunately, data were available to contrast categorization and favorability interpretations. In related pilot work, 30 behavioral items (half of which were present in the stimulus videotapes) were rated in terms of favorability and prototypicality. As expected, the prototypicality and favorability ratings were highly correlated ($r = .71, p < .001$), although this finding does not guarantee that categorization and favorability effects on leadership perceptions are identical. In fact, the correlation between the prototypicality and recognition ratings for the same 30 behaviors made by subjects in this study was significantly positive ($r = .38, p < .05$), whereas the favorability ratings were not significantly related to these recognition ratings ($r = .21, ns$). These correlational findings, as well as other work (Foti et al., 1982), suggest that categorization, rather than favorability, was the more central process in forming leadership perceptions and making behavioral ratings.

Additional concerns exist about potential reactivity of the button-pressing encoding measure. Although we assumed that subjects press the button to indicate that meaningful behavior has occurred, the subjects could signal the occurrence of a behavior via a button press, then later infer that the behavior must

be meaningful. That is, the encoding methodology may not reflect natural encoding processes but rather increase the salience of otherwise unnoticed behaviors through cued recall. Cronshaw, Staller, and Lord (1986), as part of a construct validation of the encoding measure, examined such process issues. They determined the latency between onset of individual leader behaviors presented via videotape and corresponding subject button presses, then correlated this latency measure to leadership perceptions. They found that longer latencies for encoding of prototypical leader behavior lead to stronger perceptions of leadership ($r = .30, p < .01$). Cronshaw, Staller, and Lord's findings led them to conclude that the encoding measure for prototypical behavior taps effortful cognitive processing preceding the formation of an overall leadership impression. It is important that considerable cognitive processing appeared to occur during the latency period (i.e., *before* the button presses were made). The reactivity explanation for the encoding results is therefore less plausible.

Two separate concerns about the prototypicality and consensus information manipulations should also be addressed. First, we have assumed that observers encode and process leader information using recognition-based processes. However, completion of postobservational leadership ratings may require effortful reprocessing of impression information from memory (Wyer, Srull, Gordon, & Hartwick, 1982). Therefore, leadership perceptions assessed by the pencil-and-paper measures used here may be influenced by effortful reprocessing of information occurring when the ratings were made. Second, the failure of consensus information to trigger subjects' causal ascriptions might result from the reported tendency of observers to overlook consensus data in favor of other attributional information, particularly distinctiveness (Hansen, 1980; McArthur, 1972). The theory guiding this study nevertheless predicts that consensus is important in determining leadership, and the simulated leadership situation produced conditions most likely to induce reliance on consensus information. In fact, previous research suggests that observers (as opposed to actors) prefer to base their causal attributions on consensus information (Hansen & Lowe, 1976) and that consensus information impacts to the greatest extent where it pertains to persons (McArthur, 1976). However, further research is required to establish whether other attributional manipulations (e.g., of consistency or distinctiveness) will produce an effect strong enough to override the impact of prototype information.

Managerial Implications

Our results have interesting applied implications. If organizations' participants use simplified processes to form social perceptions rather than the more careful processes suggested by attribution theory and some individuals are particularly good fits to leadership prototypes, they would be automatically recognized as leaders. This would increase their social power and imply that they were causally important in producing good outcomes. This, in turn, could affect the quality of superior-subordinate exchanges (Seers & Graen, 1984) and the credit they are given for work outcomes, and ultimately, their career paths. Alternatively, individuals who do not fit leadership prototypes may have difficulty influencing subordinates and may not be

given proper credit by superiors for good work outcomes for which they are responsible. Both types of effects may be very dysfunctional for organizations.

Similar effects may occur with other perceptual processes. For example, because they are associated with salient physical features, sex information or minority status are probably encoded using automatic, categorization-based processes. Such information could then affect expectations (Dean & Lewis, 1984) and evaluations based on subsequent information. (See Darley and Gross, 1983, for an illustration of how stereotypes and expectations can affect the use of seemingly objective information.)

Our findings also have important implications for the study of leadership. They suggest that organizational participants may often rely on simple cognitive heuristics such as categorization in forming leadership impressions, rather than using more effortful attributional reasoning. Observers may not spontaneously undertake extensive causal analyses unless faced with unexpected events (Wong & Weiner, 1981) or cued by an experimenter (Enzle & Schopflocher, 1978). Thus, researchers should direct increased attention to specifying how, under what circumstances, and what stimulus characteristics produce leadership perceptions based on categorization processes.

Our findings concerning the impact of consensus information are also interesting from a methodological perspective. Our manipulation of consensus information within the context of a realistic videotaped interaction differed considerably from previous "paper people" studies, in which subjects were explicitly informed about possible causes (e.g., Crocker et al., 1983). In the present study, subjects were required to extract causal information from on-going social interactions. When causal information is presented in this way, it may be used to a lesser extent in making subsequent judgments than previous attributional research would suggest.

In addition to exploring substantive issues, this study presented a new measurement tool for applied cognitive research. The encoding technique offers a direct behavioral measure supplementing existing pencil-and-paper and reaction time (retrieval) measures. This type of encoding technique can be applied to other important organizational issues such as performance appraisal (Banks, 1982; Kinicki, 1986) and selection interview decisions. Alternatively, the effects on leadership perceptions of verbal compared to nonverbal behavior could also be examined using this technique.

In conclusion, this study has competitively tested the impact of categorization and attribution processes on formation of leadership perceptions. Our conclusion that categorization is the primary process determining leadership perceptions has many theoretical and practical implications, but it should be replicated and extended in future research by comparing it to other attributional principles under conditions differing in information processing demands.

References

- Banks, C. G. (1982). *Cue selection and evaluation elicited during the rating process* (Working Paper No. 82-83). Austin, TX: University of Texas, Department of Management.
- Bargh, J. A., & Pietromonaco, P. (1982). Automatic information processing and social perception: The influence of trait information presented outside of conscious awareness on impression formation. *Journal of Personality and Social Psychology*, 43, 437-449.
- Bieri, J., Atkins, A. L., Briar, S., Leaman, R. L., Miller, H., & Tripodi, T. (1966). *Clinical and social judgment: The discrimination of behavioral information*. New York: Wiley.
- Calder, B. J. (1977). An attribution theory of leadership. In B. M. Staw & G. R. Salancik (Eds.), *New directions in organization behavior* (pp. 179-204). Chicago: St. Clair Press.
- Cantor, N., & Mischel, W. (1979). Prototypes in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 12, pp. 3-51). New York: Academic Press.
- Cohen, C. E. (1981). Goals and schemata in person perception: Making sense from the stream of behavior. In N. Cantor & J. F. Kihlstrom (Eds.), *Cognition, social interaction, and personality* (pp. 45-68). Hillsdale, NJ: Erlbaum.
- Crocker, J., Hannah, D. B., & Weber, R. (1983). Person memory and causal attributions. *Journal of Personality and Social Psychology*, 44, 55-66.
- Cronshaw, S. F., & Lord, R. G. (1982, February). Perceptual versus cognitive determinants of causal attributions and leadership. *Proceedings of the 25th Annual Conference of the Midwest Academy of Management* (pp. 211-220). Columbus: Ohio State University, College of Administrative Sciences.
- Cronshaw, S. F., Staller, J., & Lord, R. G. (1986). *Reliability and construct validity of an encoding measure of social perceptions*. (Working paper available from first author)
- Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, 44, 20-33.
- Dean, K., & Lewis, L. L. (1984). Structure of gender stereotypes: Interrelationships among components and gender label. *Journal of Personality and Social Psychology*, 46, 991-1004.
- DeNisi, A. S., Cafferty, T. P., & Meglino, B. M. (1984). A cognitive view of the performance appraisal process: A model and research proposition. *Organizational Behavior and Human Performance*, 33, 360-396.
- Enzle, M. E., & Schopflocher, P. (1978). Instigation of attribution processes by attributional questions. *Personality and Social Psychology Bulletin*, 4, 595-599.
- Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology*, 66, 127-148.
- Foti, R. J., Fraser, S. L., & Lord, R. G. (1982). Effects of leadership labels and prototypes on perceptions of political leaders. *Journal of Applied Psychology*, 67, 326-333.
- Foti, R. J. & Lord, R. G. (in press). Prototypes and scripts: The effects of alternative methods of processing information on rating accuracy. *Organizational Behavior and Human Decision Processes*.
- Fraser, S. L., & Lord, R. G. (1984). *The effects of stimulus prototypicality on leadership perceptions and behavioral ratings*. Unpublished manuscript, University of Akron, Akron, OH.
- Fraser, S. L., Lord, R. G., & Cronshaw, S. F. (1983). *Age and sex related differences in leadership prototypes*. Unpublished manuscript, University of Akron, Akron, OH.
- Graesser, A. C., Woll, S. B., Kowalski, D. J., & Smith, D. A. (1980). Memory for typical and atypical actions in scripted activities. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 503-515.
- Green, S. G., & Mitchell, T. R. (1979). Attributional processes of leaders in leader-member interactions. *Organizational Behavior and Human Performance*, 23, 429-458.
- Hamilton, D. L. (1979). A cognitive-attributional analysis of stereotyp-

- ing. In D. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10, pp. 53–83). New York: Academic Press.
- Hansen, R. D. (1980). Commonsense attribution. *Journal of Personality and Social Psychology*, 39, 996–1009.
- Hansen, R. D., & Lowe, C. A. (1976). Distinctiveness and consensus: The influence of behavioral information on actors' and observers' attributions. *Journal of Personality and Social Psychology*, 34, 425–433.
- Hastie, R. (1981). Schematic principles in human memory. In E. T. Higgins, C. P. Herman, & M. P. Zanna (Eds.), *Social cognition: The Ontario Symposium on Personality and Social Psychology* (pp. 39–88). Hillsdale, NJ: Erlbaum.
- Hauenstein, N. M. A., & Lord, R. G. (1986, October). *A Bayesian approach to leadership perceptions: Numbers are no substitute for experience*. Paper presented at the conference on Decision Making and Information Processing, Buffalo, NY.
- Higgins, E. T., & King, G. A. (1981). Accessibility of social constructs: Information-processing consequences of individual and contextual variability. In N. Cantor & J. F. Kihlstrom (Eds.), *Personality, cognition, and social interactions* (pp. 69–121). Hillsdale, NJ: Erlbaum.
- Higgins, E. T., Rholes, W. J., & Jones, C. R. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology*, 13, 141–154.
- Ilgel, D. R., & Feldman, J. M. (1983). Performance appraisal: A process focus. In B. M. Staw & L. Cummings (Eds.), *Research in organizational behavior* (Vol. 5, pp. 141–197). Greenwich, CT: JAI Press.
- Kelley, H. H. (1972). *Causal schemata and the attribution process*. New York: General Learning Press.
- Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, 28, 107–128.
- Kinicki, A. J. (1986). *Purpose of appraisal: The effects on rater behavior*. Manuscript submitted for publication.
- Lord, R. G. (1985). An information processing approach to social perceptions, leadership and behavioral measurement in organizations. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. 7, pp. 87–128). Greenwich, CT: JAI Press.
- Lord, R. G., & Alliger, G. M. (1985). A comparison of four information processing models of leadership and social perceptions. *Human Relations*, 38, 47–65.
- Lord, R. G., Foti, R. J., & De Vader, C. L. (1984). A test of leadership categorization theory: Internal structure, information processing, and leadership perceptions. *Organizational Behavior and Human Performance*, 34, 343–378.
- Lord, R. G., Foti, R. J., & Phillips, J. S. (1982). A theory of leadership categorization. In J. G. Hunt, U. Sekaran, & C. Schriesheim (Eds.), *Leadership: Beyond establishment views* (pp. 104–121). Carbondale: Southern Illinois University Press.
- Lord, R. G., & Smith, J. E. (1983). Theoretical, information processing, and situational factors affecting attribution theory models of organizational behavior. *Academy of Management Review*, 8, 50–60.
- Lowe, C. A., Medway, F. J., & Beers, S. E. (1976). *Individual differences in causal attribution: The Person-Environmental Causal Attribution (PECA) Scale*. Unpublished manuscript, University of Connecticut, Storrs, CT.
- McArthur, L. A. (1972). The how and what of why: Some determinants and consequences of causal attribution. *Journal of Personality and Social Psychology*, 22, 171–193.
- McArthur, L. Z. (1976). The lesser influence of consensus than distinctiveness information on causal attributions: A test of the person-thing hypothesis. *Journal of Personality and Social Psychology*, 33, 733–742.
- Markus, H., & Smith, J. (1981). The influence of self-schemata on the perception of others. In N. Cantor & J. F. Kihlstrom (Eds.), *Personality, cognition, and social interaction* (pp. 233–262). Hillsdale, NJ: Erlbaum.
- Newton, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, 28, 28–38.
- Newton, D., Engquist, G., & Bois, J. (1976). The reliability of a measure of behavior perception. *Psychological Documents* (formerly JSAS: Catalog of Selected Documents in Psychology), 6, 5. (Ms. No. 1173)
- Newton, D., Engquist, G., & Bois, J. (1977). The objective basis of behavior units. *Journal of Personality and Social Psychology*, 35, 847–862.
- Newton, D., Rindner, R., Miller, R., & LaCross, K. (1978). Effects of availability of feature changes on behavior segmentation. *Journal of Experimental Social Psychology*, 14, 379–383.
- Overall, J. E., & Woodward, J. A. (1977). Nonrandom assignment and the analysis of covariance. *Psychological Bulletin*, 84, 588–594.
- Pfeffer, J. (1977). The ambiguity of leadership. *Academy of Management Review*, 2, 104–112.
- Phillips, J. S. (1984). The accuracy of leadership ratings: A cognitive categorization perspective. *Organizational Behavior and Human Performance*, 33, 125–138.
- Phillips, J. S., & Lord, R. G. (1981). Causal attributions and perceptions of leadership. *Organizational Behavior and Human Performance*, 28, 143–163.
- Phillips, J. S., & Lord, R. G. (1982). Schematic information processing and perceptions of leadership in problem-solving groups. *Journal of Applied Psychology*, 67, 486–492.
- Rush, M. C., Phillips, J. S., & Lord, R. G. (1981). Effects of a temporal delay in rating on leader behavior descriptions: A laboratory investigation. *Journal of Applied Psychology*, 66, 442–450.
- Seers, A., & Graen, G. B. (1984). The dual attachment concept: A longitudinal investigation of the combination of task characteristics and leader-member exchanges. *Organizational Behavior and Human Performance*, 33, 283–306.
- Smith, E. R., & Miller, F. D. (1983). Mediation among attributional inferences and comprehension processes: Initial findings and a general method. *Journal of Personality and Social Psychology*, 44, 492–505.
- Srull, T. K., & Wyer, R. S., Jr. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology*, 37, 1660–1672.
- Stein, R. T., & Heller, T. (1979). An empirical analysis of the correlations between leadership status and participation rates reported in the literature. *Journal of Personality and Social Psychology*, 37, 1993–2002.
- Wilder, D. A. (1981). Perceiving persons as a group: Categorization and intergroup relations. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 213–257). Hillsdale, NJ: Erlbaum.
- Wong, P. T. P., & Weiner, B. (1981). When people ask “why” questions, and the heuristics of attributional search. *Journal of Personality and Social Psychology*, 40, 650–663.
- Wyer, R. S., Srull, T. K., Gordon, S. E., & Hartwick, J. (1982). Effects of processing objectives on the recall of prose materials. *Journal of Personality and Social Psychology*, 43, 674–688.

Received February 24, 1986 ■

Task Planning and Energy Expended: Exploration of How Goals Influence Performance

P. Christopher Earley
Claremont McKenna College and
Claremont Graduate School

Pauline Wojnaroski
Claremont Graduate School

William Prest
Pitzer College

Although many studies demonstrate the effectiveness of goal setting in organizations, it is unclear how goals actually influence performance. In present studies we examined the effect of assigned goals and task information on performance, energy expended, and task planning or organizing. In Study 1, a 2×2 (Information \times Goal) design was used. Subjects were 72 undergraduates working on a business simulation. In Study 2, 129 male and female workers from a service organization and a moving company responded to a survey assessing an individual's goal setting, job training, energy expended during a typical task performance, and task planning undertaken prior to performance. The results of both studies demonstrated that goal setting and task training influenced the dependent variables. In addition to influencing an individual's energy expended (effort and persistence), having a specific goal led an individual to plan and organize more than an individual given a general goal (i.e., "do your best"). The results of both studies suggest that goal setting and task-relevant information influence performance, in part, through their influence on energy expended and planning.

The utility of assigning individuals specific and challenging work goals has been well documented (Earley, 1985; Latham & Steele, 1983; Locke, Shaw, Saari, & Latham, 1981). The specific mechanisms through which assigned goals influence performance can be placed in two categories (Locke et al., 1981), those having a direct effect (effort, persistence, and directed attention) on an individual and those having an indirect effect (strategy development). An important distinction between these two types of effects is that the direct influence is primarily motivational, that is to say, an allocation of the individual's energy-related resources to task performance, whereas the indirect effect is primarily cognitive, that is, the way in which a plan or strategy is developed to use the mobilized, energy-related resources. In the present study we examine the direct and indirect effects of goals on performance, and explore the general importance of a strategy, or plan, in the relation of goals to performance. For the purpose of the present research, direct effects (energy-related resources) were limited to the effort and persis-

tence components described by Locke et al., (1981). The indirect effect of a goal was defined as the strategy development, or planning, engaged in by the individual.

The direct form of goal influence on an individual's behavior has been explored in several studies. LaPorte and Nath (1976) found that subjects provided with difficult goals spent more time (persistence) on reading passages than did subjects with easy goals. Sales (1970) and Bassett (1979) found that people with difficult performance objectives tried harder on tasks than did people with easy ones.

A possible reason for the effect of specific goals on an individual's mobilization of energy-related resources (energy expended) has been suggested by several researchers. Bandura and Schunk (1981) found that subjects with specific (proximal) goals developed a greater sense of task mastery than did subjects with general (distal) goals. They concluded that the sense of mastery enhanced an individual's intrinsic interest in the task, and that a person who is interested in a task will try harder. Specific goals may also have a beneficial effect on effort because the goal makes clear what is expected of the individual. Role ambiguity has been found to have emotionally disruptive consequences (Earley & Kanfer, 1985; Lee & Schuler, 1982; Rakestraw & Weiss, 1981).

The planning mechanism of goal setting refers to the procedure (sequence of behaviors) used by an individual to translate his or her resources into action (Schank & Abelson, 1977). The influence of specific goals on an individual's planning has not been widely tested, but several studies suggest the effect of goals on planning. Terborg (1976) found that subjects who set specific goals spent more time on a task using relevant task strategies (e.g., note-taking to learn prose passages). Rosswork (1977)

The authors' research was supported in part by a grant to the first author from the Claremont McKenna College Research and Travel Fund.

The authors would like to thank Dwight Richards and Peter Johnson for their assistance in collecting the field data. The authors would also like to thank C. Lee, E. A. Locke, and the anonymous reviewers for their helpful comments on an earlier draft of this article. We would also like to thank F. Landy for his helpful editorial guidance during the review process.

Correspondence concerning this article should be addressed to P. Christopher Earley, who is now a faculty member in the Department of Management, University of Arizona, Tucson, Arizona 85721.

found that subjects who were given a difficult goal wrote short sentences for a task requiring sentence writing. Huber (1985) found that difficult goals led to less effective plan development (strategy) than did easy goals for a heuristic task. Locke et al., (1981), and Smith, Locke, and Barry, (1985), found that assigning goals to individuals led them to develop high quality plans to achieve their goals. Using business students in an organizational simulation, Smith et al. (1985) found that formal planning was unrelated to performance, but that the quality of planning was related to performance. Most of these studies did not attempt to assess planning as a function of the goals set (see Huber, 1985; Smith et al., 1985, for exceptions).

The effect of specific goals on an individual's planning is primarily one of information processing. After receiving a specific goal, an individual is directed to the task and must decide how to proceed (Campbell, 1984b; Earley, 1986; Earley & Kanfer, 1985) prior to performing. A specific goal contains more information than a general goal, namely, the precise level of performance expected of the individual. A specific goal stimulates the development of task-relevant plans by stimulating an individual to think more about a task; after receiving a specific goal an individual must decide how to achieve that particular performance level and how to use his or her personal resources in doing so. Someone with a general goal does not have a specific performance level about which to think, and therefore spends less time thinking about how to work on the task.

The previously described research suggests that goal setting influences performance in at least two ways—in the energy expended and in the planning done by the individual. Therefore, it was hypothesized that an individual assigned a specific goal would expend more effort, plan more, and outperform an individual given a general goal. It was also hypothesized that energy expended and planning would mediate the relation of goal setting to performance.

The amount of planning an individual engages in on a particular task is influenced by the information available about the task in the work environment. A specific goal represents one type of information. Many other types of information are available to an individual in a work setting. An additional important form of information available to an individual is the knowledge he or she has concerning the task itself. If a goal stimulates planning by providing an individual with task-related information, then it should be possible to influence planning by providing a worker with other types of information, including job knowledge (Earley, 1986). Locke, Frederick, Lee, and Bobko (1984) found that training individuals in specific task strategies influenced their actual strategies and, thus, their performance. Earley (1985, 1986) found similar results for presenting individuals with performance strategies and job context (information about the task). These studies suggest that planning is influenced by the information provided to an individual. Providing an individual with information about a task (e.g., the important aspects of a task) should stimulate planning as does assigning a specific goal. Therefore, it was hypothesized that providing an individual with task-relevant information would increase planning and that increased planning would result in increased performance. To the extent that planning involves the use of energy-related resources, an individual with task-relevant infor-

mation should also mobilize more energy during task performance. Therefore, it was also hypothesized that individuals provided with task-relevant information would expend more effort while working on a task than would individuals lacking this information.

In summary, the present study proposes that planning and energy expended mediate the relation of goal setting to performance and the relation of task-relevant information to performance. It is also proposed that planning, energy expended, and performance will be higher for those individuals having a specific goal or a great deal of task knowledge than for those having a general goal or little task knowledge. To test these hypotheses, two studies were conducted. Study 1, a laboratory investigation, was intended to assess the relations among the variables under highly controlled conditions. Study 2 was a field survey intended to broaden and generalize the results of the laboratory study to a variety of jobs.

Study 1

Method

Subjects. A total of 72 undergraduates from a west-coast college participated in the study as partial fulfillment of a class requirement.

Design. A 2×2 (Goal \times Information) crossed, factorial design was used. The goal manipulation consisted of a subject either receiving an assigned goal or instructed to "do your best." Information consisted of a subject either receiving specific information about the task (high) or not receiving this information (low).

Task and goal. The task required a subject to read a paragraph describing a fictitious product, to choose a medium in which to advertise it (from a list of eight possible media), and to write statements in support of his or her choice of a medium. The products ranged from household goods to business computers. The goal assigned to the subjects in the assigned goal condition was to produce at least four arguments justifying their medium choice for each of 35 products in 60 min (a total of 140 arguments). During a pilot study, 4% of the subjects achieved this goal. The subjects in the do-your-best condition were instructed to "do their best" while working on the task.

Information manipulation. Although all of the subjects were provided basic instructions concerning the task, subjects in the high-information conditions were provided a *media fact sheet* containing several considerations to be made when choosing an advertising medium for a product. The media fact sheet emphasized that consideration of who will buy the product, where it will be distributed, when it will be advertised, and how much money is available for advertising, would be useful. The media fact sheet presented a brief description of each of these considerations.

A confederate posed as an advertising expert who would provide the subjects information (via the media fact sheet) that they might find helpful as they worked on the task. The fact sheet consisted of a two-page outline describing four major considerations to be made when advertising a product. The four points were (a) the buying population, (b) locations for product distribution, (c) the timing of the ads, and (d) the advertising budget and cost considerations. By providing this information, it was believed that the subjects would structure their strategy (develop a plan for writing justifications) using the information provided.

Measures. The planning or organizing activities of an individual were assessed using the following four items:

1. "Please describe in as much detail as possible the methods and/or framework you used to try to achieve the goal," coded as the total num-

Table 1
Means and Standard Deviations for Measures of Performance,
Planning, and Energy Expended for Goal
and Information in Study 1

Measure and information	Goal			
	"Do best"		Specific	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Performance				
High	8.26	1.98	20.25	6.72
Low	15.32	2.26	25.14	6.56
Planning				
High	1.78	0.49	3.23	0.52
Low	2.57	0.73	3.39	0.85
Energy expended				
High	2.61	0.60	3.24	0.63
Low	3.35	0.53	3.76	0.59

Note. *n* = 18/cell.

ber of separate steps listed (the total number of steps listed ranged from 1 to 7).

2. "Organizing a procedure for this task was _____ for accomplishing as much as I did," wherein 1 = *not at all important* and 5 = *extremely important*.

3. "To work on the task I _____," wherein 1 = *just started working* and 5 = *carefully planned how to work toward my goal by considering time pressures and other elements*.

4. "To accomplish the goal assigned to me I was _____ about the manner in which I tried to finish the whole task," wherein 1 = *not at all deliberate* and 5 = *extremely deliberate*.

The reliability (Cronbach's alpha) of these four items was .76. For subsequent analyses, a composite score was formed for planning using the mean of the responses to these items.

Energy expended (effort and persistence) was assessed by the following three items:

1. "I worked at this task without getting tired for _____," wherein 1 = *only a very short time* and 5 = *a very long time*.

2. "How much effort did you expend to work on this task?" wherein 1 = *little or no effort* and 5 = *almost all the effort I could*.

3. "While working on the task, I found myself working _____," wherein 1 = *not at all hard* and 5 = *extremely hard*.

The reliability (Cronbach's alpha) of these three items was .73. For subsequent analyses, a composite score was formed for energy expended using the mean of the responses to these items.

The planning and energy-expended items were factor analyzed using a principal-components analysis and a varimax rotation. The data were analyzed using a two-factor solution that accounted for 64% of the total variance. The energy-expended items loaded on the first factor (loadings ranged between .43 and .71), and the planning items loaded on the second factor (loadings ranged between .53 and .79). The cross loadings for both sets of items ranged between .03 and .11. Although the intercorrelation between the composite scores for planning and energy expended was moderate ($r = .33$), the results of the factor analysis suggest that the items assessed separable constructs.

Procedure. Upon entering the experimental room, subjects were asked to be seated at individual desks containing the experimental materials. Next, the subjects were asked to read the provided task instructions as they were read aloud by the experimenter and to practice on a sample item for 3 min. They were also asked if they had any questions as the experimenter picked-up any completed materials.

In the high-information manipulation, the subjects were introduced to a confederate posing as an advertising expert from a local university. The expert distributed the media fact sheets (described earlier) and read them aloud to the subjects. Next, the expert asked if the subjects had questions, and she instructed the subjects to write down the four basic advertising criteria on the products' packet sitting on their desk. Finally, the expert instructed the subjects that they had a few minutes before the task would begin. The subjects in the low-information conditions were not provided any specific information, but were addressed by the confederate on an irrelevant topic (current school social events) for an equivalent amount of time as the subjects in the high-information manipulation.

After the information manipulation was enacted, the subjects were either assigned the goal or told to "do your best." The subjects were asked to complete a preperformance questionnaire assessing perceived task difficulty. After completing the questionnaire, the subjects were instructed to begin working on the task.

When subjects completed the 60-min performance period, all of their completed materials were picked up and they were given a few minutes to rest. Next, the experimenter administered a questionnaire assessing the manipulation checks, the subject's planning and energy expended on the task. Finally, the subjects were debriefed and thanked.

Results

Manipulation checks. Three items were used to assess the information manipulation: "How useful was any information you may have been provided while working on the task?", "How much information did you get from the experimenter concerning how to actually write arguments?", and "How much information did you receive about the task and how to achieve your goal from the experimenter?" Each was rated on a 5-point scale wherein lower scores reflect lower amounts of information conveyed. The reliability (Cronbach's alpha) of these items was .83. A two-way analysis of variance (ANOVA) using the goal-setting and information conditions demonstrated a significant main effect for information, $F(1, 68) = 175.41, p < .01, M_{\text{high}} = 3.88, M_{\text{low}} = 1.75$. No other significant effects were obtained.

The goal manipulation was assessed by two items: "How specific was the goal you were given?" wherein 1 = *not at all specific* and 5 = *extremely specific*, and, "How detailed was your goal?" wherein 1 = *not at all detailed* and 5 = *extremely detailed*. The correlation between these items was .91. A two-way ANOVA using the goal-setting and information conditions conducted on a composite score for these items demonstrated a significant main effect for goal, $F(1, 68) = 37.03, p < .01, M_{\text{assigned}} = 4.04, M_{\text{do your best}} = 2.28$. No other significant effects were obtained.

Table 2
Pearson Correlations for Measures in Study 1

Measure	1	2	3	4	5
1. Performance ^a	—	.56	.51	.68	.39
2. Planning ^a		—	.33	.25	.63
3. Energy expended ^a			—	.45	.37
4. Goal ^b				—	.00
5. Information ^b					—

^a All correlations are significant at $p < .05$.

^b Dummy coded 0, 1.

Table 3
Hierarchical Regression Analyses for Study 1

Variable	Step	R^2	R^2 change (per step)	β	T (for beta)
3A. Performance					
Energy expended	1	.45	.45	.34	3.21*
Planning				.42	3.95*
Goal setting	2	.69	.24	.57	7.28*
Information				.15	1.67 _{ns}
3B. Planning					
Goal setting	1	.46	.46	.25	2.95*
Information				.63	7.11*
3C. Energy expended					
Goal setting	1	.35	.35	.45	4.67*
Information				.37	3.85*

* $p < .05$.

Finally, checks for differences in practice trial performance (ability) and perceived task difficulty demonstrated no significant differences across the experimental conditions.

Performance. The means and standard deviations, performance, planning, and energy expended are presented in Table 1. Correlations for these measures are presented in Table 2.

The performance score for each subject was derived from the total number of arguments produced during the 60-min performance period. (Two assistants, blind to the experimental conditions, coded the number of arguments produced. The interrater correlation was .89 ($p < .01$)). A two-way ANOVA was conducted to test the hypotheses that goal setting and information provision would significantly influence performance, and to assess differences across the experimental conditions.

The analyses demonstrated a significant main effect for goal and information, $F(1, 68) = 88.03, 26.41, p < .05$, for the goal and information main effects, respectively. As hypothesized, individuals who received high information outperformed those receiving low information. Further, individuals who received a specific goal outperformed individuals who received a general goal. No significant interaction was obtained.

Regression analyses. In order to test the hypotheses that planning and energy expended would mediate the effects of goal setting and information provision on performance, a series of hierarchical regression analyses was conducted to construct a path model of the data, using a procedure described by Pedhazur (1982, pp. 577–633). The specific order of the variables was based on a priori considerations, as well as on the temporal ordering of the variables. The results of these analyses are presented in Table 3.

The first analysis (3A) was a hierarchical regression conducted on performance using planning, energy expended (Step 1), and goal setting (dummy coded 0,1 for the “do your best” and specific conditions, respectively) and information (dummy coded 0,1 for the low- and high-information conditions, respectively; Step 2). Additional analyses were conducted on planning (3B) and energy expended (3C), using goal setting and information provision. The results of these analyses are summarized in Figure 1A.

The adequacy of the proposed model was tested using the procedure outlined by Pedhazur (1982, pp. 618–622). Specifically, a goodness-of-fit statistic W that is distributed as chi-square is calculated based on an index of fit, Q ($Q = .998$). The results demonstrate that the overidentified model in Figure 1A adequately describes the obtained data, $\chi^2(1, N = 72) = .198, p > .05$.

The results of the analyses failed to support the hypothesis that planning and energy expended would mediate the effect of goals on performance, but these two variables were found to have a partial mediating role in the relation of information provision to performance. The results also supported the hypothesis that specific goals would lead to greater performance, planning, and energy expended during task performance compared to general goals. Likewise, the hypothesis that providing individuals with high task-relevant knowledge would increase their performances, planning, and energy expended more than those not provided with knowledge, was supported.

Summary. The results of the laboratory study support the hypotheses that goals and information influence planning and energy expended on a task. It was also demonstrated that both planning and energy expended significantly predicted performance, and that the goal manipulation significantly predicted performance even after the variance due to planning and energy had been statistically controlled. Thus, goals appear to influence performance through some unmeasured variables as well as through planning and energy expended. These unmeasured variables may include effort or direction, but may not have been fully assessed with the perceptual measures used in the present study. The information manipulation, however, appears to influence performance solely through planning and energy expended. (A regression entering information prior to planning and energy demonstrates that information predicts 22% of the variance in performance. After removing the variance due to planning and energy expended, information no longer accounted for a significant proportion of unique variance in performance. Refer to James and Brett, 1984, for a further discussion of the procedure used to test a mediational hypothesis).

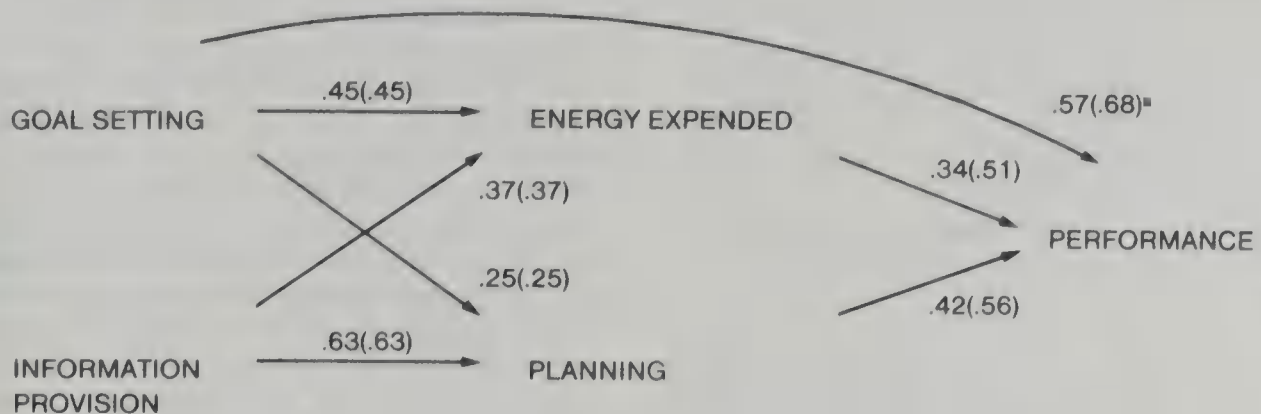
Study 2

The results of Study 1 suggest that planning and energy expended play a mediating role in the relation of information to performance, and a partial mediating role in the relation of goals to performance. A second study was conducted to further explore the relations suggested in Study 1. This study consisted of a field survey assessing the relations among the variables for the general work patterns of workers. Rather than focusing on a particular task, the purpose of Study 2 was to assess the general relation between an individual's amount of goal setting across a variety of tasks, job training received, typical planning performed, typical energy expended, and a supervisor's performance rating of the individual. Thus, Study 2 extends Study 1 to a range of tasks performed by a worker in a more naturalistic and dynamic setting.

Method

Subjects. A survey was distributed to 155 male and female workers from a service organization and a moving company (approximately one

(1A) Laboratory Study



(1B) Field Study

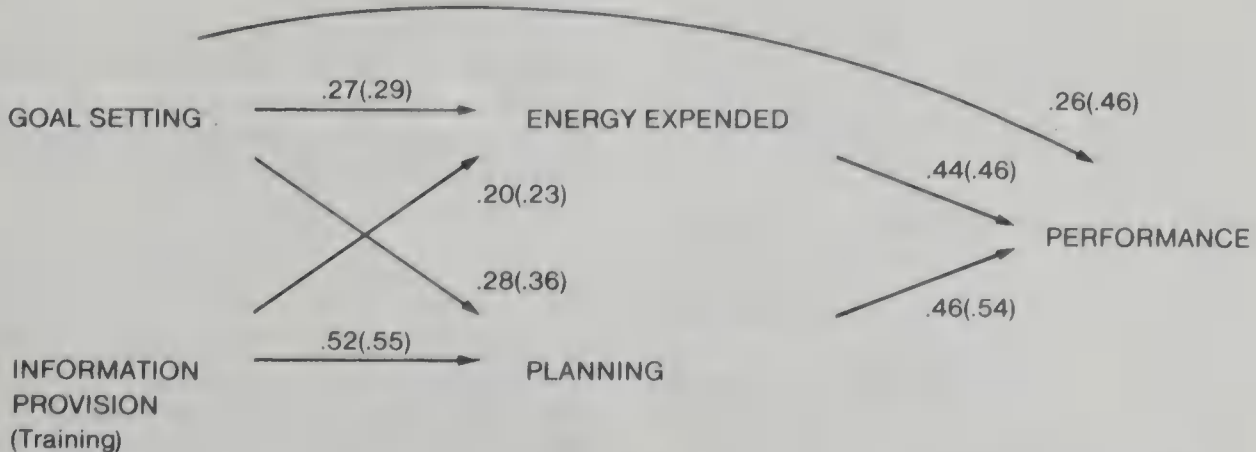


Figure 1. General summary models depicting the influence of goals and task-relevant information (training) on planning, energy expended, and performance. (*Zero-order correlations, shown in parentheses, are significant at $p < .05$, along with standardized path coefficients.)

half from each) at the beginning of a work shift. The workers were given a work break in which to fill out the surveys. A total of 129 surveys were completed and returned, for a return rate of 83%. The means for age and tenure at their present jobs were 36.14 and 7.37 years, respectively. The employees came from a variety of job categories (13% production, 20% clerical, 49% service, and 17% management).

Survey instrument. The survey consisted of items assessing biographical data, energy expended and planning performed while working on a typical work activity, amount of goal setting, and prior task training. The items used were adapted from those used in Study 1. In addition, each worker's supervisor was asked to provide performance ratings for the survey respondents. After matching the performance ratings with the questionnaires, the respondents' names were removed from the questionnaires.

Specifically, performance was assessed using the following two items: "How would you rate this worker's typical performance on a project?" and "How would you rate the performance of this worker on the last three projects (or tasks) he or she worked on?" Anchors were *extremely poor* (1) and *extremely good* (5). The correlation between these items

was .91 ($p < .05$). For subsequent analyses, a composite score was formed for performance, using the mean of the responses to these items. Energy expended was measured using the following three items:

1. "How much effort do you exert while working on a typical project?" wherein 1 = *very little effort* and 5 = *alot of effort*.
2. "While working on my day-to-day tasks, I find myself working _____," wherein 1 = *not at all hard* and 5 = *extremely hard*.
3. "While working on a project, I find myself _____," wherein 1 = *always distracted by other projects* and 5 = *never distracted by other projects*.

The reliability (Cronbach's alpha) of these items was .81 ($p < .05$). For subsequent analyses, a composite score was formed for energy expended using the mean of the responses to these items. Planning was assessed using the following three items:

1. "Please describe step-by-step how you went about working on your last work project," coded as the total number of separate steps listed (the number of steps ranged from 1 to 8).
2. "Which statement best describes your approach to a new project

Table 4
Descriptive Statistics and Correlations in Study 2

Measure	<i>M</i>	<i>SD</i>	1	2	3	4	5
1. Performance	3.58	0.89	—	.54	.46	.46	.31
2. Planning	4.56	1.29		—	.29	.36	.55
3. Energy expended	3.30	0.77			—	.29	.23
4. Goal	3.50	1.33				—	.16
5. Training	3.26	1.45					—

Note. *n* = 129. All correlations are significant at *p* < .05.

or task?” wherein 1 = *I dig right in and get started right away* and 5 = *I think a long time about how to approach the project, then I proceed*.

3. “Organizing an elaborate procedure or scheme is typically _____ for accomplishing a project,” wherein 1 = *not at all important* and 5 = *extremely important*.

The reliability (Cronbach) of these three items was .79 (*p* < .05). For subsequent analyses, a composite score was formed for planning using the mean of the responses to these items. As with Study 1, the planning and energy-expended items were factor analyzed using a principal components analysis (with a varimax rotation). The results of the two-factor solution demonstrate that the planning items loaded on the first factor (loadings ranged between .58 and .81) and the energy-expended items loaded on the second factor (loadings ranged between .50 and .63). The cross loadings for each set of items ranged between .07 and .21. The intercorrelation between the composite scores for planning and energy expended was moderate (*r* = .29).

An individual’s amount of specific goal setting was assessed using the following two items:

1. “How would you characterize your own work objectives for performance,” wherein 1 = *my goals are general* (e.g., “do my best”) and 5 = *my goals are specific* (e.g., make 20 sales calls).

2. “How often do you set specific goals for your own, daily work? (e.g., to make 20 sales calls each day)” wherein 1 = *not at all often* and 5 = *extremely often*.

(The example used in these items “to make 20 sales calls” was varied according to the job of the employee. For instance, a furniture mover was provided the example “make 15 delivery stops.” The level of difficulty implied by each example was held constant based on the opinion of job incumbents who participated from the various positions during a pilot study.) The correlation between these items was .73 (*p* < .05).

Prior job training was assessed using the following three items:

1. and 2. “How much training did you receive for your present position? From a formal training program . . . From an informal training program (e.g., watching others, word-of-mouth),” wherein 1 = *very little training* and 5 = *a lot of training*.

3. “How useful is any information you may have received during training to your work performance?” wherein 1 = *not at all useful* and 5 = *extremely useful*.

The reliability (Cronbach’s alpha) of these items was .86 (*p* < .05).

Results

Descriptive statistics and correlations. The means, standard deviations, and zero-order correlations for performance, energy expended, planning, goals, and prior training are presented in Table 4.

Regression analyses. A set of analyses similar to that used in Study 1 was conducted for Study 2. The results of these analyses are presented in Table 5 and are summarized in Figure 1B.

The adequacy of the proposed model was tested using Pedhazur’s (1982, pp. 618–622) procedure. The results demonstrate that the overidentified model in Figure 1B adequately describes the obtained data, $\chi^2(1, N = 129) = 3.06, Q = .938, p > .05$.

The analyses provide support for the hypothesis that planning and energy expended mediate the effect of training on performance, but fail to support a mediational role of planning and energy expended in the relation of goal setting to performance. In addition, the analyses demonstrate support for the hypotheses that performance, planning, and energy expended would be higher for those individuals having had a lot of training or having specific goals than for those having had little training or having general goals, respectively.

Summary. The analyses for both Studies 1 and 2 suggest that goals and training influence performance, in part, by stimulating an individual’s amount of planning and level of energy expended. Further, the results support the hypotheses that task-relevant information and goals influence both planning and energy expended, which in turn influence performance. Goals also have a direct influence on performance, not fully described by planning and energy expended. Finally, the analyses suggest that both goals and training influence the amount of planning an individual engages in. Therefore, a cognitive benefit of goals is to get an individual to think about both the task and how to perform it. The relations among these variables were assessed for typical tasks performed by the respondents, and the results represent a general pattern for work behavior.

Discussion

An important finding of the present studies is that goals influence performance through a variety of mechanisms including the following: increasing energy expended (a person’s effort and persistence) and increasing a person’s planning or organizing for a task. In addition, goals influence performance in a fashion similar to other types of job information such as job knowledge.

Investigators have suggested that goals influence performance through several mechanisms. The indirect influence of goals on planning has been found to be especially important.

Table 5
Hierarchical Regression Analyses for Performance for Study 2

Dependent variable	Step	<i>R</i> ²	<i>R</i> ² change (per step)	β	<i>T</i> (for beta)
Performance					
Energy expended	1	.39	.39	.44	4.56*
Planning				.46	6.12*
Goal setting	2	.44	.05	.26	3.07*
Training				.04	0.25
Planning					
Goal setting	1	.39	.39	.28	3.92*
Training				.52	7.10*
Energy expended					
Goal setting	1	.12	.12	.27	3.07*
Training				.20	2.46*

* *p* < .05.

Planning, however, may not always prove to be beneficial. For instance, Smith et al., (1985) found that formal planning opportunity is not as important as the quality of planning. In their study, managers were required to meet with one another in a formalized planning session prior to the outset of each of the subsequent simulation games. They found that the quantity of planning (time spent) did not predict performance. In the present studies an individual's amount of planning (number of steps in a given plan) was related to his or her task performance. The planning engaged in and the energy expended by an individual, however, did not fully account for all of the effects of goal setting on performance. This finding may be partly attributable to the preliminary nature of the planning measure. The specific measure used may not adequately reflect planning, although it appeared to be face valid based on reactions of the respondents during a pilot test of the survey instrument. Perhaps a multidimensional measure of planning (including quality and breadth) would better account for the relation of goals to performance.

Goal setting also had an effect on the amount of energy expended by an individual (increasing an individual's effort and persistence). Effort expended alone, however, did not fully explain the variance in an individual's performance; consideration of an individual's plan was also important in describing the effects of goals on an individual's performance. The challenge represented by the goal (Locke et al., 1981) induced higher rated levels of effort, but without an appropriate plan the added effort was not effective in increasing an individual's performance. This conclusion is supported by Campbell (1984a), who in a post hoc analysis found that individuals lacking familiarity with a complex task did not benefit by having a specific goal.

The effect on performance of providing an individual with a general understanding of his or her job can be explained by the intervening mechanisms of increased planning and effort. Providing individuals with job-relevant information stimulates both their planning activities and the way they approach their tasks. The effect of training information on energy expended may operate as the result of increasing an individual's self-efficacy expectations (Bandura & Schunk, 1981; Earley, 1985; Locke et al., 1984). People who have high self-efficacy expectations will continue working hard on a task even if the task (or goal) is extremely difficult.

The similarity of the effects of training and goals on planning and energy expended suggests an important general mechanism in task performance. The processing of goal information and other forms of task information stimulates an individual's level of planning. Although planning as used in the present study did not fully explain the relation of goals to performance, it is clearly an important element in the goal setting model. Task information produced effects similar to those resulting from the assignment of a goal; thus, it appears that the translation of work information (e.g., goals, task knowledge) to performance plans is a key in determining how and why goals influence performance. This is not to say that the energy-expended variables are unimportant in goal setting, but planning represents another relevant issue.

In the present studies, planning and energy expended were assumed to be independent mediators in the goal to perfor-

mance relation. The obtained correlations between these constructs, however, were moderate in magnitude, suggesting some connection between them. One plausible explanation for such a relation may be derived by examining the role of attention in the processing of goal information. Locke et al., (1981) suggested that goals "most fundamentally" direct an individual's attention to what must be done. Goals and task information direct an individual's attention, thereby stimulating planning. Inasmuch as attention can be viewed as a processing capacity, or allocatable resource (Kahneman, 1973), it is likely that goal-directed attention functions as an energizer of the planning process. Task difficulty influences the amount of attention allocated by an individual (Kahneman, 1973). In this sense, the influence of goals attributable to directed attention motivates planning. Therefore, goals may influence an individual in the following manner: directed attention (motivational, or resource allocation effect) → plan development → energy expended in accordance to plan (motivational effect). This model suggests the cyclic nature of motivation and cognition and warrants future research.

Several limitations of the present study should be addressed in considering the findings. As suggested earlier, the items assessing planning may reflect characteristics other than just planning. For instance, an item used in Study 2 ("Please describe step-by-step how you went about working _____") may measure an individual's ability for self-expression in addition to actual planning. Further, the perceptual measure of energy expended may not accurately reflect an individual's effort because of an individual's introspective limits (Locke et al., 1981). Finally, the path analyses are suggestive of the relations among the variables, but are not necessarily causal and should not be interpreted as such (Pedhazur, 1982, pp. 577-580). Thus, although the results of the two studies complement one another and are provocative, they are preliminary and merit further exploration before firm conclusions can be reached.

The present study has shown beneficial effects for goals and training on performance, energy expended, and planning or organizing. It was found that assigning goals to individuals directs their attention to the task at hand; having a lengthy performance strategy, however, does not necessarily imply that an individual will perform at a high level. Future work should be directed at developing a more thorough measure of individual-level plans and determining the process through which plans develop.

References

- Bandura, A., & Schunk, D. H. (1981). Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. *Journal of Personality and Social Psychology*, 41, 586-598.
- Bassett, G. A. (1979). A study of the effects of task goal and schedule choice on work performance. *Organizational Behavior and Human Performance*, 24, 202-227.
- Campbell, D. J. (1984a). The effect of goal-contingent payment on the performance of a complex task. *Personnel Psychology*, 37, 23-40.
- Campbell, D. J. (1984b). *Task complexity and strategy development: A review and conceptual analysis*. Manuscript submitted for publication.
- Earley, P. C. (1985). The influence of information, choice, and task complexity upon goal acceptance, performance, and personal goals. *Journal of Applied Psychology*, 70, 481-491.

- Earley, P. C. (1986). Supervisors and shop stewards as sources of contextual information in goal setting: A comparison of the U.S. with England. *Journal of Applied Psychology*, 71, 111-118.
- Earley, P. C., & Kanfer, R. (1985). The influence of component participation and role models on goal acceptance, goal satisfaction, and performance. *Organizational Behavior and Human Decision Processes*, 36, 378-390.
- Huber, V. L. (1985). Effects of task difficulty, goal setting, and strategy on performance of a heuristic task. *Journal of Applied Psychology*, 70, 492-504.
- James, L. R., & Brett, J. M. (1984). Mediators, moderators, and tests for mediation. *Journal of Applied Psychology*, 69, 307-321.
- Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice Hall.
- LaPorte, R. E., & Nath, R. (1976). Role of performance goals in prose learning. *Journal of Educational Psychology*, 68, 260-264.
- Latham, G. P., & Steele, T. P. (1983). The motivational effects of participative versus assigned goal setting on performance. *Academy of Management Journal*, 26, 406-417.
- Lee, C., & Schuler, R. S. (1982). A constructive replication and extension of a role and expectancy perception model of participation in decision making. *Journal of Occupational Psychology*, 55, 109-118.
- Locke, E. A., Frederick, E., Lee, C., & Bobko, P. (1984). Effect of self-efficacy, goals, and task strategies on task performance. *Journal of Applied Psychology*, 69, 241-251.
- Locke, E. A., Shaw, K. N., Saari, L. M., Latham, G. P. (1981). Goal setting and task performance: 1969-1980. *Psychological Bulletin*, 90, 125-152.
- Pedhazur, E. J. (1982). *Multiple regression in behavioral research*. New York: Holt, Rinehart & Winston.
- Rakestraw, T. L., & Weiss, H. M. (1981). The interaction of social influences and task experience on goals, performance, and performance satisfaction. *Organizational Behavior and Human Performance*, 27, 326-344.
- Rosswork, S. G. (1977). Goal setting: The effects on an academic task with varying magnitudes of incentive. *Journal of Educational Psychology*, 69, 710-715.
- Sales, S. M. (1970). Some effects of role overload and role underload. *Organizational Behavior and Human Performance*, 5, 592-608.
- Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding*. Hillsdale, NJ: Erlbaum.
- Smith, K. G., Locke, E. A., & Barry, D. (1985). *The relationship between planning, goal setting and organizational performance: An experimental study*. Manuscript submitted for publication.
- Terborg, J. R. (1976). The motivational components of goal setting. *Journal of Applied Psychology*, 61, 613-621.

Received March 11, 1986

Revision received May 5, 1986 ■

Patterns of Work and Nonwork Satisfaction

Garnett Stokes Shaffer
University of Georgia

The relation between work and nonwork satisfaction has been the subject of much research, but results have been inconclusive. The present study explored the possibility that different satisfaction profiles exist for different people, accounting for the past inconsistent findings. Varying patterns of satisfaction were identified by subgrouping individuals on the basis of their work satisfaction and nonwork satisfaction profiles. Items measuring work and nonwork satisfaction were factored for 398 female and 390 male college graduates using a principal components analysis; a hierarchical subgrouping was conducted based on subjects' profiles on the satisfaction components. A number of satisfaction profiles were identified. Further analyses indicated that background experiences and current work and nonwork activities were associated with the satisfaction types. Implications for future research are discussed.

In spite of beliefs in a substantial relation between work and life outside of work, most research demonstrates only a weak empirical relation between work and nonwork satisfaction (Near, Rice, & Hunt, 1980; Rice, Near, & Hunt, 1980). Several reasons have been suggested. One is that methodological weaknesses contribute to the inconsistent and frequently weak findings. However, some methodological problems would lead to the conclusion that the results are inflated. The present study addresses a different possibility: that the important role of individual differences in work and nonwork attitudes has been overlooked.

The notion that individual differences moderate the covariation of work and nonwork satisfaction is not new. Sex (Kavanagh & Halpern, 1977; Near et al., 1980), job importance or focus (Dubin, 1956; Iris & Barrett, 1972), personality (Bedeian & Marbert, 1979; Champoux, 1978, 1981; Schmitt & Bedeian, 1982), job level (Kavanagh & Halpern, 1977), and education and income (Bamundo & Kopelman, 1980) have all been related to the relation between work and nonwork satisfaction. However, like moderator research in other domains, the findings are often inconsistent.

The weak, inconsistent findings are evidence that a general law about the work-nonwork satisfaction relation may not exist (Kabanoff & O'Brien, 1980). However, the range of correlations is wide, suggesting that there may be groups of individuals with different satisfaction profiles.

Rather than search for groups of individuals who demonstrate a different relation between work and nonwork, that is, search for groups with positive, negative, and no correlation between work and nonwork, a more fruitful approach would be

to identify first what work and nonwork satisfaction patterns are common among workers.

Only Kabanoff and O'Brien (1980) divided people into groups with different work-nonwork profiles; however, they examined only one aspect of nonwork activity—leisure. Moreover, satisfaction was not studied; instead, their focus was on attributes of work and leisure behaviors. In spite of their more limited focus, their results provide some impetus for an investigation of groups of individuals with different satisfaction profiles. A promising avenue for such an investigation is the use of empirical subgrouping. The identification of groups of individuals with different satisfaction profiles—that is, different satisfaction types—may help explain some of the inconsistency in the literature. In addition, the presence of subgroups argues for the possibility of seeking lawful relations that apply to homogeneous groups of people and provides a new avenue for future research on work and nonwork attitudes.

Subgroup Differences

The identification of homogeneous subgroups, each with similar profiles of work and nonwork attitudes, is not very meaningful in the absence of differences across the groups on other psychologically important variables. Some personal characteristics and situational variables may be more important for developing some satisfaction profiles than for others. For example, Schneider (1985) suggested that a particular level of pay that may be very satisfying to a worker in mental health may be very dissatisfying to someone who is in business. If variables such as pay are uniquely associated with some subgroups of individuals who have internally similar profiles, but not with others, then this is evidence that the subgroups moderate the relations. Three possible types of variables are examined in the present study: background, work-related, and nonwork-related variables.

Background Variables

Friend and Haggard (1948) and others (Near et al., 1980; Stiles, 1985) have noted the importance of family history in

The author would like to thank two anonymous reviewers for their comments and suggestions.

Correspondence concerning this article should be addressed to Garnett Stokes Shaffer, University of Georgia, Department of Psychology, Athens, Georgia 30602.

determining later satisfaction. Other background influences include socioeconomic status (Inkeles, 1960), social activities (Champoux, 1981), and sports participation (Varca, Shaffer, & Saunders, 1984). Perhaps socioeconomic status would differentiate between group members whose dissatisfaction is based on a discrepancy between current status and adolescent status and group members whose dissatisfaction is associated with other variables. Previous involvement in leisure might differentiate between groups who are satisfied with work and nonwork in general and those who are not. A poor family history might predict membership in groups that are unhappy with both work and nonwork relations.

Work-Related Variables

Both career success and intentions to leave a job have been related to satisfaction (Bray, Campbell, & Grant, 1974; Bray & Howard, 1980; Lounsbury, Gordon, Bergermaier, & Francesco, 1982). Occupational success may differentiate between groups with high pay satisfaction who are highly paid and those who are not highly paid.

Nonwork-Related Variables

The role of leisure activities in determining work and nonwork satisfaction is unclear (Haavio-Mannila, 1971; Kabanoff, 1980; Kabanoff & O'Brien, 1980). Some researchers argued that leisure activities may be selected to compensate for activities not experienced at work. It may be that leisure activities are only important to the satisfaction of individuals who had active lives growing up. The same may be true for the role of social orientation and activities. A specific leisure activity associated with satisfaction is religious participation and involvement (Vecchio, 1980). For some groups, participation in religious activities may be correlated with work and nonwork satisfaction, but low religious participation should not necessarily be related to dissatisfaction for all groups.

Reliance on Bivariate Models

Researchers studying work and nonwork satisfaction have tended to rely on either global measures of satisfaction or summary measures of work and nonwork facets. Just as the job satisfaction literature makes clear that work satisfaction is multivariate and complex (Scarpello & Campbell, 1983), it is also probable that nonwork satisfaction is multifaceted. The reliance on a bivariate model of work and nonwork satisfaction probably has obscured a number of interesting relations, such as those noted by Chacko (1983). The present study goes beyond the traditional bivariate definitions of work and nonwork satisfaction to represent their full complexity.

Summary of Present Study

The present study is exploratory and its purpose is twofold. First, the presence of subgroups of individuals with internally similar and externally dissimilar profiles of work and nonwork satisfaction is investigated to identify satisfaction types. Work and nonwork satisfaction are examined as multidimensional at-

titudes. Second, specific background, work-related, and nonwork-related factors should vary across types, or subgroups, indicating the complexity of the relations with many variables previously associated with work and nonwork satisfaction. Positive findings would provide evidence that the patterns of work and nonwork satisfaction of workers are more complex than the previous research has indicated, and consequently, changes implemented to increase the quality of life will not have the same effects on all workers.

Method

Subjects

Subjects in the present investigation were 398 female and 390 male college graduates who were freshmen at the University of Georgia in 1968 or 1970 and who completed a Post College Experience Inventory (PCEI) in 1980. The questionnaire was used as part of a larger longitudinal investigation described by Owens and Schoenfeldt (1979).

Questionnaires

Work and nonwork satisfaction. Items measuring work and nonwork satisfaction were obtained from the Post College Experience Inventory, developed by Owens and Schoenfeldt (1979), and mailed to alumni of the University of Georgia in 1980, either 6 or 8 years after college graduation. The PCEI contains 97 items about job activities, job-seeking behavior, leisure activities, religious involvement, and other adult life experiences. Approximately 70% of those sent the PCEI returned it, and a study conducted by Jackson (1978) indicated no significant differences in the backgrounds of respondents and nonrespondents.

Background variables. As freshmen at the University of Georgia, each member of the sample had completed a Biographical Questionnaire (BQ; Owens & Schoenfeldt, 1979), an 118-item questionnaire tapping the general life experiences of adolescents. Items included were questions about parental and sibling relationships, socioeconomic status, religious activities, athletic interests, scientific interests and activities, and a range of other feelings, activities, and interests. The BQ had been previously factored using a principal components analysis with a varimax rotation. Fifteen female and 13 male factors were judged interpretable, and component scores were available for each individual in the sample. Based on the examination of previous literature, several factors were included as background variables in the present study. A 5-year retest reliability estimate of the scores derived from the factors indicated that each of the factors used in the present study had a reliability coefficient above .75 (Shaffer, Saunders, & Owens, 1986).

Component scores (i.e., responses weighted by loadings on the principal components) were used as background factors:

1. Parental relationships: For men, a factor named "warmth of parental relationship" was used. The factor measures closeness to parents, parental interest in activities, and emotional support from parents. For women, two factors represented parental relationship: "warmth of maternal relationship" and "warmth of paternal relationship."
2. Socioeconomic status: This factor indicates the social class, income, and education of parents.
3. Sports participation: This factor measures the amount of interest, involvement, and competence in athletic pursuits.
4. Social activities: For men, a factor labeled "social introversion" was used. The items tapped participation in social activities, parties, and dating activity. For women, a factor labeled "popularity with the opposite sex" tapped involvement in social and dating activities.

Work and nonwork variables. The PCEI included items about adult

Table 1
Rotated Component Loading Matrix for Work and Nonwork Satisfaction Items

Item	Men				Women			
	Job	Relationships and leisure activities	Environment	Job relationships	Pay	Intrinsic job elements	Personal and social relationships	Environment
Nature of work itself	.72	.08	.20	.32	-.12	.81	.14	.11
Importance of job in life	.69	-.04	-.07	-.06	-.04	.51	-.15	-.05
If could start over, would go into same line of work ^a	.67	.20	-.06	-.07	.19	.65	.16	-.02
Plan to remain in current line of work ^b	.66	.02	.15	.17	.24	.68	.03	.06
Opportunity for individual discretion and responsibility	.62	-.01	.10	.45	-.10	.70	.13	-.10
Opportunity for advancement	.59	.02	.04	.35	.40	.43	.04	.07
Reputation of company and its management	.56	-.05	.06	.41	.21	.38	-.02	.01
Intellectual stimulation	.07	.83	.00	.00	.03	.09	.80	.00
Companionship	.00	.82	-.10	.08	.06	.00	.83	.09
Social activities	.14	.77	.28	.06	.05	.03	.79	.24
Recreational and leisure activities	.09	.72	.25	-.01	-.02	.01	.78	.20
Size of immediate family	-.15	.61	.12	.28	.12	.03	.47	.22
Friendship acquaintances	.04	.42	.52	.33	-.03	.12	.31	.60
Community	.13	.14	.83	.10	.04	.04	.12	.75
Living quarters	-.02	.08	.72	.05	.32	-.03	.09	.77
Relations with fellow workers	.06	.12	.18	.70	.02	.55	.01	.38
Relations with supervisor	.16	.11	.04	.67	.03	.48	.07	.00
Working conditions	.15	.06	.00	.60	.18	.43	-.06	.02
Compensation and benefits	.14	.02	.04	.29	.78	.06	-.01	.17
Standard of living	.13	.18	.37	-.10	.72	-.10	.29	.62

Note. All items are scored on a 5-point continuum from very satisfied to very dissatisfied unless otherwise noted.
^a Scored on a 3-point continuum.
^b Scored on a 5-point continuum from plan to remain indefinitely to do not plan to remain.

experiences. Based on previous research on the determinants of work and nonwork satisfaction, several indices of work-related and nonwork-related factors were developed from PCEI items. For the work-related variables they included the following:

- 1. Occupational success: Two items asked about monthly income and the number of salary increases since college graduation. The income item is limited in variability, particularly for men, because the highest income option is for a gross monthly income above \$1,200.
- 2. Turnover intentions: A single item index asked "Which of the following best describes your intention of staying with your present company or organization?" Four options were provided, ranging from *plan to stay indefinitely* to *do not intend to stay*.

The nonwork related variables included the following:

- 1. Religious involvement: Two items tapped this dimension, one item asking about frequency of weekly church service attendance and the second item asking about frequency of midweek service attendance.
- 2. Social orientation: Four items measured social orientation, including such items as number of friends of the same and opposite sex, and participation in social clubs and in civic or professional organizations.
- 3. Leisure activities: Three kinds of items were used to measure leisure activities, including amount of spare time spent on hobbies (one item), sports (one item), and reading (three items).

Results

Work and Nonwork Satisfaction Factors

Twenty items from the PCEI were used as measures of work and nonwork satisfaction. Rather than simply summing the

items to form a single work factor and a single nonwork factor, the items were factored using a principal components analysis and the matrix was rotated to an orthogonal criterion solution. This provided an opportunity to examine more detailed profiles among various facet measures. An orthogonal solution was selected to meet the requirements of the clustering procedure. Analyses were conducted separately for men and women.

For men, five factors were selected and rotated, accounting for 60% of the total variance. For women, four factors were selected and rotated, accounting for 55% of the total variance. The factors are named and described in Table 1, and a full matrix of factor loadings is provided. For men, the factors included two nonwork factors and three work factors. Two work and two nonwork factors emerged for the women.

Subgrouping on Satisfaction Profiles

To subgroup individuals based on their profiles, component scores were computed, each weighted by factor loadings and standardized with a mean of 0 and standard deviation of 1. The subjects' component scores were converted to a matrix of inter-subject profile similarity, a D^2 matrix with each cell representing the profiles of a pair of subjects. The Ward and Hook (1963) hierarchical procedure was applied to the subgrouping of the profiles. The Ward and Hook procedure forms clusters so that within-group variation is minimized and between-group variation is maximized at each stage. The number of subgroups was

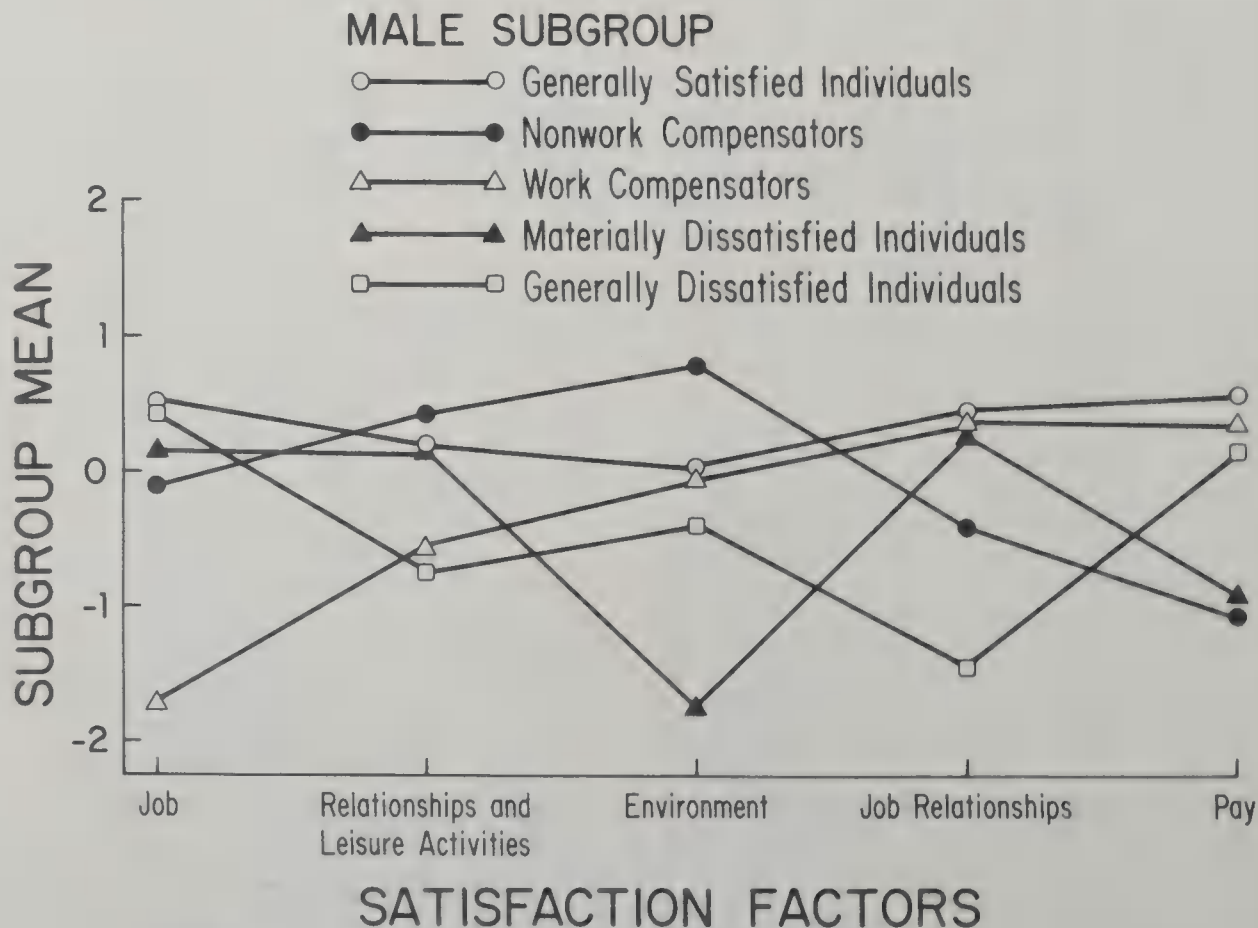


Figure 1. Male subgroup profiles on five work and nonwork satisfaction factors.

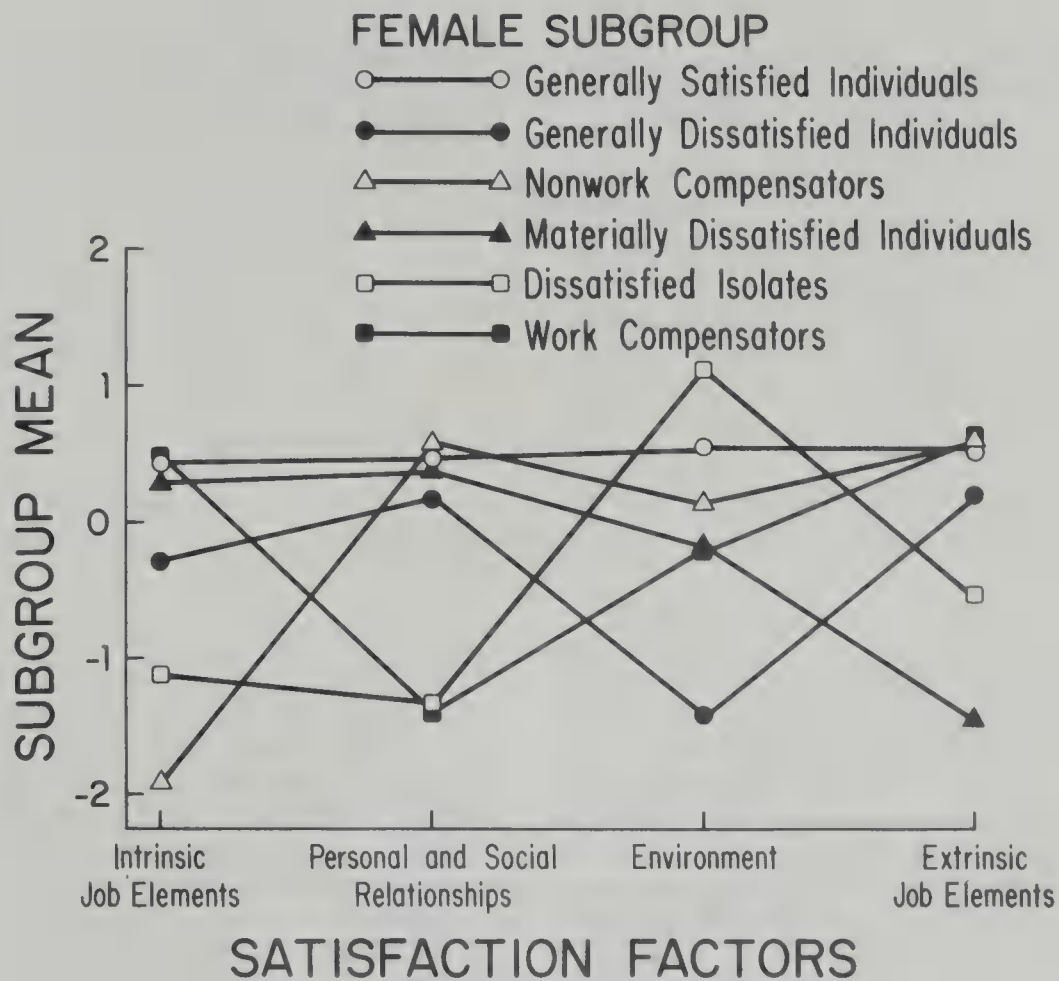


Figure 2. Female subgroup profiles on four work and nonwork satisfaction factors.

determined by examining the increment in within-group variation that occurs as the number of subgroups is reduced to one. Examination of a plot of the incremental within-groups sums of squares often points to one or more solutions. When this occurred, each potential solution was examined and the final number was determined based on the solution interpretability, the size of the groups, and the number of misfits, that is, individuals with profiles not matching any one subgroup profile. Fol-

lowing subgroup selection, group membership for each subject was reaffirmed, and if necessary, modified, inasmuch as the iterative nature of the clustering procedure can result in a group mean shift (Feild & Schoenfeldt, 1975).

Five subgroups were selected for men, and six groups were selected for women. Approximately 9% of each sample (men and women) were identified as misfits. Mean subgroup profiles on the work and nonwork satisfaction factors are described in

Table 2
Subgroup Factor Comparisons, Analysis of Variance, and Mean Differences for Men

Group	M satisfaction with				
	Job	Relationships and leisure activities	Environment	Job relationships	Pay
1	0.45 ^a	0.16 ^{a,b}	0.10 ^b	0.42 ^a	0.44 ^a
2	-0.18 ^b	0.48 ^a	0.78 ^a	-0.45 ^b	-0.96 ^c
3	-1.69 ^c	-0.52 ^c	-0.04 ^{b,c}	0.35 ^a	0.33 ^{a,b}
4	0.25 ^a	0.09 ^b	-1.72 ^d	0.18 ^a	-0.85 ^c
5	0.39 ^a	-0.88 ^d	-0.32 ^c	-1.40 ^c	0.12 ^b
<i>F</i>	103.94	23.06	63.37	59.22	52.34
ω^2	.51	.18	.39	.37	.35

Note. Means for each factor = 0, $SD = 1$. Any groups having at least one common superscript are *not* significantly different on that factor. All F s are significant at $p < .0001$.

Table 5
Intercorrelation Matrix for Dependent Variables for Women

Variable	1a	1b	2	3	4	5	6	7	8	9	10	11
Background												
1. Parental relationship												
a. Maternal	—	.31	-.22	-.14	.11	-.10	-.03	.05	-.09	.07	-.07	.02
b. Paternal		—	-.07	-.09	.12	-.10	-.04	-.07	-.15	.03	.00	-.01
2. Socioeconomic status			—	.18	.16	.12	.13	-.06	.19	-.01	.11	.09
3. Sports participation				—	.20	-.06	-.05	-.17	-.08	-.13	.19	-.09
4. Social activities					—	.03	.05	.09	.17	-.07	.01	-.10
Work												
5. Occupational success						—	.27	-.25	.00	.00	.06	-.01
6. Turnover intentions							—	-.03	.01	.01	.05	-.05
Nonwork												
7. Religious involvement								—	.16	.08	-.18	-.08
8. Social orientation									—	.15	.14	.17
Leisure activity												
9. Hobbies										—	.07	.07
10. Sports											—	.07
11. Reading												—

Table 6
Means Across Subgroups for Significant Background, Work, and Nonwork Factors

Factor	Generally satisfied individuals	Nonwork compensators	Work compensators	Materially dissatisfied individuals	Generally dissatisfied individuals
Male subgroups					
Background ^a					
Parental relations	515 ^a	492 ^{ab}	487 ^{ab}	497 ^{ab}	456 ^b
Social activities ^b	490 ^b	512 ^{ab}	534 ^{ab}	527 ^{ab}	554 ^a
Socioeconomic status	479 ^a	504 ^a	499 ^a	515 ^a	520 ^a
Sports participation	517 ^a	491 ^{ab}	465 ^b	484 ^{ab}	505 ^{ab}
Work related ^c					
Turnover intentions	3.69 ^a	2.76 ^b	2.20 ^c	2.64 ^b	2.93 ^b
Occupational success	8.70 ^a	7.73 ^b	7.82 ^{ab}	8.09 ^{ab}	8.13 ^{ab}
Nonwork related ^c					
Social orientation	10.69 ^a	10.35 ^a	9.12 ^b	8.68 ^b	9.24 ^b
Reading	9.10 ^b	9.70 ^b	9.60 ^b	9.73 ^b	10.96 ^a
Female subgroups					
	Generally satisfied individuals	Generally dissatisfied individuals	Nonwork compensators	Materially dissatisfied individuals	Dissatisfied isolates
Background ^a					
Parental relations					
Maternal	521 ^a	485 ^a	507 ^a	494 ^a	480 ^a
Paternal	501 ^{ab}	524 ^a	496 ^{ab}	498 ^{ab}	458 ^{ab}
Socioeconomic status	538 ^a	483 ^b	527 ^{ab}	497 ^{ab}	508 ^{ab}
Work related ^c					
Turnover intentions	3.11 ^a	2.88 ^a	1.63 ^c	2.34 ^b	2.26 ^b
Occupational success	7.33 ^{abc}	7.50 ^{ab}	6.93 ^c	6.33 ^c	6.56 ^{bc}
Nonwork related ^c					
Social orientation	10.76 ^a	8.90 ^b	9.48 ^a	9.65 ^b	9.35 ^b
Hobbies	2.63 ^b	2.33 ^b	3.22 ^a	2.58 ^b	2.65 ^b

Note. All means with the same superscripts are not significantly different ($p < .05$).
^a Measured from the Biographical Questionnaire. Factors have mean of 500 and standard deviation of 100.
^b High scores indicate low social participation.
^c Measured from Post College Experience Inventory.

Although the apparent variation in the subgroup profiles suggested the presence of different subgroups, the differences were tested using one-way analyses of variance (ANOVAS), as suggested by Schoenfeldt (1969). Subgroup membership served as the independent variable in each analysis, and the dependent variables were the work and nonwork satisfaction factor scores. Tables 2 and 3 contain the results of the analyses, all of which are highly significant. Subgroup mean differences were tested using the Newman-Keuls test. Clearly, the individual groups differed in terms of their satisfaction profiles.

Differences in Subgroup Profiles

After identification of subgroup profiles, the next question was whether the groups differed in terms of background and current situations. The data were analyzed using ANOVAS, with subgroup membership as the independent variable and each of the background, work, and nonwork indices as separate dependent variables. The intercorrelations among the dependent variables are reported in Tables 4 and 5. Significant ANOVAS were further analyzed by examining mean differences using the Newman-Keuls test. Only individuals who were classified as good fits in the subgroup analysis were used.

Background variables. Results of the ANOVAS indicated that several factors varied across subgroup membership. Parental relationships varied across groups for both men and women: warmth of parental relationship (men), $F(4, 345) = 4.22, p < .01$; warmth of maternal relationships (women), $F(5, 357) = 2.68, p < .05$; and warmth of parental relationship (women), $F(5, 357) = 2.24, p < .05$. An examination of the subgroup means provided in Table 6 indicates that for men, the generally satisfied group had better parental relationships than did the generally dissatisfied group. No significant mean differences were obtained for the female groups on the maternal relationship factor, but the generally dissatisfied women reported significantly closer paternal relationships than did the dissatisfied isolate women.

Socioeconomic status (SES) also varied across subgroups for both men and women, $F(4, 345) = 2.49, p < .05$, and, $F(5, 357) = 3.85, p < .01$, respectively. No significant mean differences were found for men on the SES factor. However, the generally dissatisfied women came from a significantly lower SES background than did the generally satisfied group.

Sports participation and social activities varied significantly across only the male subgroups, $F(4, 345) = 3.35, p < .05$, and, $F(4, 345) = 4.58, p < .01$, respectively. The post hoc tests indicated that the generally satisfied group participated in more athletic activities in high school than did the work compensators and was more socially active in high school than the generally dissatisfied group.

Work-related variables. As expected, occupational success and turnover intentions varied significantly across subgroups. For men, significant differences were obtained for occupational success, $F(4, 345) = 4.96, p < .001$, and turnover intentions, $F(4, 345) = 32.03, p < .0001$. The generally satisfied group was more financially successful than the nonwork compensators and was more likely to plan to stay with current work organizations than any other group. The work compensators were more

likely than any other group to report plans to leave their jobs, possibly because of their dissatisfaction with the nature of their work. This finding may seem to contradict the naming of the subgroup, but the impetus for the name was that the group was satisfied with job relationships and dissatisfied with nonwork relationships.

Significant differences on the work-related variables were also obtained for the female groups: occupational success, $F(5, 357) = 6.90, p < .0001$, and turnover intentions, $F(5, 357) = 13.26, p < .0001$. The work compensators reported more financial success and less desire to leave their jobs than did all of the other groups. The materially dissatisfied women reported the least success financially, but the nonwork compensators were most likely to indicate plans to leave their jobs.

Nonwork-related variables. For men, the social orientation and reading variables varied across patterns of subgroup membership, $F(4, 345) = 8.26, p < .0001$, and, $F(4, 345) = 2.71, p < .05$. Both the generally satisfied individuals and the nonwork compensators were significantly higher than the other groups in terms of their involvement in social and civic clubs and their number of friends. The generally dissatisfied individuals were significantly higher than the other groups in terms of their reading activity.

Among the women, two variables varied significantly across subgroup membership, social orientation, $F(5, 357) = 6.68, p < .0001$, and hobbies, $F(5, 357) = 2.99, p < .05$. The nonwork compensators reported spending significantly more spare time with hobbies than did the other groups, and along with the generally satisfied groups were higher in their social involvement than the other groups.

Discussion

Note that many of the moderators of work and nonwork satisfaction were controlled to some extent in this study. Men and women were analyzed separately, and the sample was fairly homogeneous in terms of education, current SES, and career stage because sample members were all college graduates who completed bachelor's degrees 6 to 8 years prior to reporting work and nonwork satisfaction. Yet the presence of various profiles in a relatively homogeneous population suggests that contradictory findings in previous studies regarding the direction of the relation between work and nonwork satisfaction may have occurred because different relations exist for different subgroups of individuals. The fact that work, nonwork, and background variables significantly differed across the subgroups is an indication that the groups themselves have some psychological meaningfulness.

Frequently, a low, positive relation between work and nonwork satisfaction has been found (Near et al., 1980). This can be understood in light of findings in the present study. The largest subgroup for both men and women was the generally satisfied individuals. The presence of such a group in any study, along with the presence of a fairly large generally dissatisfied group, would lead to a low, positive relation between work and nonwork. However, groups with highly variable profiles, such as work and nonwork compensators, also represented a significant proportion of the sample. In these groups, a compensatory, or

negative, relation between work and nonwork was suggested. The presence of each of these subgroups provides evidence that the world of attitudes is more complex than previous research has demonstrated.

Findings of individual differences frequently lead to questions about the influence organizations can have on the satisfaction of their workers. However, from both the researcher's and the organization's perspectives, the presence of subgroups offers reasons for optimism. Some relationships may be applicable to some groups of individuals, providing generalizability that is more limited than universal laws, which would apply to all of the workers, but more promising than the possibility that there are no laws that can be applied to any worker. Subgrouping individuals to identify homogeneous patterns of experience has been argued for and successfully applied in other contexts (Owens & Schoenfeldt, 1979). Such an approach offers a personal component to our conceptions of work and nonwork that has been ignored until recently (Schneider, 1985; Staw & Ross, 1985). Further research should examine the situational or personal determinants of subgroup types.

The present results also indicate that a bivariate model of work and nonwork satisfaction is too simplistic for explaining the nature of the relationship between work and life outside of work. However, the question of the number and nature of work and nonwork satisfaction factors and subgroups needs to be investigated further. The present study was a fairly conservative test of the number of subgroups, in that the sample was a relatively homogeneous one compared to the world of work at large. The number of subgroups may be greater in a more heterogeneous sample.

Work and nonwork satisfaction may also significantly change over time (Bray et al., 1974). Van Maanen and Katz (1976) have argued that change must be central to any account of a person's relationship to a job. Bray and Howard (1980) discovered patterns of satisfaction among managers that differed from those found in the present study, suggesting that profiles of satisfaction may not be stable across the adult years. Their sample consisted of individuals who were older and well established in their careers, whereas the present sample was relatively young and not yet established. Future research should examine shifts in patterns and determine what factors lead to the development of various satisfaction patterns.

The present research was exploratory. The patterns and relationships identified were based on a homogeneous group of college graduates. All data were obtained from self-report questionnaires, and much of it was attitudinal rather than behavioral. No global measure of work, nonwork, or life satisfaction was available, making it impossible to examine satisfaction patterns at a global level. Scarpello and Campbell (1983) noted that global measures of job satisfaction are tapping more than the sum of facet measures, and a recent study (Stiles, 1985) indicated that global measures may be important in understanding the relation between work and nonwork satisfaction. Future research should include both facet and global measures.

References

- Bamundo, P. J., & Kopelman, R. E. (1980). The moderating effects of occupation, age, and urbanization on the relationship between job satisfaction and life satisfaction. *Journal of Vocational Behavior*, 17, 106-123.
- Bedeian, A. G., & Marbert, L. D. (1979). Individual differences in self-perception and the job-life satisfaction relationship. *Journal of Social Psychology*, 109, 111-118.
- Bray, D. W., Campbell, R. J., & Grant, D. L. (1974). *Formative years in business: A longterm AT&T study of managerial lives*. New York: Wiley.
- Bray, D. W., & Howard, A. (1980). Career success and life satisfaction of middle-aged managers. In L. A. Bond & J. C. Rosen (Eds.), *Competence and coping during adulthood*, (pp. 258-287). Hanover, NH: University Press of New England.
- Chacko, T. I. (1983). Job and life satisfaction: A causal analysis of their relationships. *Academy of Management Journal*, 26, 163-169.
- Champoux, J. E. (1978). Work, central life interests, and self-concept. *Pacific Sociological Review*, 21, 209-220.
- Champoux, J. E. (1981). An exploratory study of the role of job scope, need for achievement, and social status in the relationship between work and nonwork. *Sociology and Social Research*, 65, 153-176.
- Dubin, R. (1956). Industrial workers' worlds: A study on the central life interests of industrial workers. *Social Problems*, 4, 131-142.
- Feild, H. S., & Schoenfeldt, L. F. (1975). Ward and Hook revisited: A two-part procedure for overcoming a deficiency in the grouping of persons. *Educational and Psychological Measurement*, 35, 171-173.
- Friend, J. G., & Haggard, E. A. (1948). *Work adjustment in relation to family background* (Applied Psychological Monograph No. 16). Stanford, CA: Stanford University Press.
- Haavio-Mannila, E. (1971). Satisfaction with family, work, leisure, and life among men and women. *Human Relations*, 24, 585-601.
- Inkeles, A. (1960). Industrial man: The relation of status to experience, perception, and value. *American Journal of Sociology*, 66, 1-31.
- Iris, B., & Barrett, G. V. (1972). Some relations between job and life satisfaction and job importance. *Journal of Applied Psychology*, 56, 301-304.
- Jackson, K. E. (1978). *Characteristics of responders and nonresponders to a series of mail surveys: A longitudinal study*. Unpublished masters' thesis, University of Georgia.
- Kabanoff, B. (1980). Work and nonwork: A review of models, methods, and findings. *Journal of Applied Psychology*, 64, 596-609.
- Kabanoff, B., & O'Brien, G. L. (1980). Work and leisure: A task attributes analysis. *Journal of Applied Psychology*, 65, 596-609.
- Kavanagh, M. J., & Halpern, M. (1977). The impact of job level and sex differences on the relationship between life and job satisfaction. *Academy of Management Journal*, 20, 66-73.
- Lounsbury, J. W., Gordon, S. R., Bergermaier, R. L., & Francesco, A. M. (1982). Work and nonwork sources of satisfaction in relation to employee intention to turnover. *Journal of Leisure Research*, 14, 285-294.
- Near, J. P., Rice, R. W., & Hunt, R. G. (1980). The relationship between work and nonwork domains: A review of empirical research. *Academy of Management Review*, 5, 415-429.
- Owens, W. A., & Schoenfeldt, L. F. (1979). Toward a classification of persons. *Journal of Applied Psychology*, 64, 569-597.
- Rice, R. W., Near, J. P., & Hunt, R. G. (1980). The job satisfaction/life satisfaction relationship: A review of empirical research. *Basic and Applied Social Psychology*, 1, 37-64.
- Scarpello, V. G., & Campbell, J. P. (1983). Job satisfaction: Are all the parts there? *Personnel Psychology*, 36, 577-600.
- Schneider, B. (1985, August). *The people make the place*. Presidential address presented at the Society for Industrial and Organizational Psychology at the meeting of the American Psychological Association, Los Angeles.
- Schoenfeldt, L. F. (1969, April). *Methodological considerations in the*

- subgrouping of persons*. Paper presented at the meeting of the Southern Society for Philosophy and Psychology, Miami, Florida.
- Schmitt, N., & Bedeian, A. G. (1982). A comparison of LISREL and two-staged least squares analysis of a hypothesized life-job satisfaction reciprocal relationship. *Journal of Applied Psychology*, 67, 806-817.
- Shaffer, G. S., Saunders, V., & Owens, W. A. (1986). *Further evidence for the accuracy of biographical data: Long-term retest and observer reports*. Manuscript submitted for publication.
- Staw, B. M., & Ross, J. (1985). Stability in the midst of change: A dispositional approach to job attitudes. *Journal of Applied Psychology*, 70, 469-480.
- Stiles, D. M. (1985). *Family background as an antecedent of job satisfaction*. Unpublished doctoral dissertation, University of Georgia.
- Van Maanen, J., & Katz, R. (1976). Individuals and their careers: Some temporal considerations for work satisfaction. *Personnel Psychology*, 29, 601-616.
- Varca, P. E., Shaffer, G. S., & Saunders, V. (1984). A longitudinal investigation of sport participation and life satisfaction. *Journal of Sport Psychology*, 6, 440-447.
- Vecchio, R. P. (1980). A test of the job satisfaction-job quality relationship: The case of religious affiliation. *Journal of Applied Psychology*, 65, 195-201.
- Ward, J. H., & Hook, N. E. (1963). Application of an hierarchical grouping procedure to the problem of grouping profiles. *Educational and Psychological Measurement*, 23, 69-81.

Received January 21, 1986

Revision received August 26, 1986 ■

Searches Open for Editors of Five APA Journals

The Publications and Communications Board is seeking nominations for new editors for the following five APA journals: *Journal of Applied Psychology*, *Journal of Comparative Psychology*, *Journal of Experimental Psychology: Human Perception and Performance*, *Professional Psychology: Research and Practice*, and *Psychological Review*. All terms will run from 1989 to 1994.

Candidates must be members of APA and should be available to start receiving manuscripts in January 1988 to prepare for issues published in 1989. Each nomination should be accompanied by a supporting statement of one page or less and should be submitted by February 15, 1987 to the Chair of the Search Committee for the relevant journal, as specified below. Please note that the P&C Board encourages more participation by women and ethnic minority men and women in the publication process, and would particularly welcome such nominees.

Journal of Applied Psychology—Robert Guion is the incumbent editor. Nominations should be submitted to Kay Deaux, Center for Advanced Studies in the Behavioral Sciences, 202 Junipero Serra Boulevard, Stanford, California 94305. The other members of the search committee are Daniel Ilgen, Raymond Katzell, Robert Ramos, and Elizabeth Loftus.

Journal of Comparative Psychology—Jerry Hirsch is the incumbent editor. Nominations should be submitted to Martha Storandt, Department of Psychology, Washington University, St. Louis, Missouri 63130. The other members of the search committee are Irwin Bernstein, David Chiszar, George Collier, and Bennett Galef.

Journal of Experimental Psychology: Human Perception and Performance—William Epstein is the incumbent editor. Nominations should be submitted to Anne Danielson Pick, ATTENTION: Journal Search Committee, Institute of Child Development, University of Minnesota, Minneapolis, Minnesota 55455. The other members of the search committee are Walter Gogel, Steven Keele, and Anne Triesman.

Professional Psychology: Research and Practice—Norman Abeles is the incumbent editor. Nominations should be submitted to Barbara Strudler Wallston, Box 512, Peabody MRL Room 415, Vanderbilt University, Nashville, Tennessee 37203. The other members of the search committee are Allan Barclay, Douglas Bray, Donald Freedheim, Sandra Haber, Florence Kaslow, Joseph Matarazzo, Carolyn Payton, and Richard Suinn.

Psychological Review—Martin Hoffman is the incumbent editor. Nominations should be submitted to Frances Horowitz, Human Development, University of Kansas, 130 Haworth, Lawrence, Kansas 66045. The other members of the search committee are Richard Nisbett and Michael Posner.

Experimental Test of an Emotion-Based Approach to Fitting Brand Names to Products

Albert Mehrabian and Robert de Wetter
University of California, Los Angeles

A theoretically derived approach to fitting brand names to products was proposed and tested with university undergraduates as subjects. The basic premise of the approach was that any product conveys a wide range of connotations to consumers and that product appeal can be enhanced by selecting a name that conveys a desirable subset of those connotations. A comprehensive, three-dimensional system for measuring emotional states was used. Using the three emotion scales, one group of subjects rated how they ideally would wish to feel while using a given product. A second group used the emotion scales to rate the emotions connoted by each of several names assigned to that product. Results of these two sets of ratings for each product and each name yielded discrepancy scores (on the dimensions of pleasure, arousal, and dominance) between the ideally desired emotional impact of the product and the emotional impact of the selected name. We hypothesized that the latter discrepancy scores are negative correlates of preference for products assigned specific names. To test the hypothesis, a third group of subjects provided ratings of preference (liking, desire to purchase) for various product-name combinations. Results, based on a highly reliable measure of product preference, supported the hypothesis. Discrepancy scores accounted for 30% of variance in product preferences of males and 37% of variance for females, thus providing strong support for the proposed theoretical model that fits brand names to products.

The study presented in this article dealt with the problem of fitting brand names to products. The few research-oriented studies available focused on recall of a brand name or brand mark as an important dependent measure of the success of either. For example, Block (1969) investigated recall and correct association of corporate symbols (or marks) with the corporations. He found descriptive symbols or marks to be superior to nondescriptive marks in helping consumers recall company names and their types of business. Block concluded that "it might be wise for a company to develop a name which can be represented symbolically in a suggestive sort of way" (p. 412).

Kanungo (1968) categorized brand names according to high versus low meaningfulness (e.g., *Legs* as high meaningful versus *Leget* as low meaningful for nylons) and fitting versus nonfitting (e.g., *Letters* as fitting, and *Economy* as nonfitting for writing pad). Recall of the brand names was the dependent measure. Consistent with the findings of Block (1969) with corporate marks, recall was superior for fitting, compared with nonfitting, brand names; it also was superior for high-meaningful, compared with low-meaningful, brand names.

Lutz and Lutz (1977) also dealt with meaningfulness of associations between brand or company names on the one hand and products and services on the other. *Interactive* images were defined as those that integrated parts of both the brand (or company) name with parts of the product (or service) in a single, integrated picture. An example of this was *Rocket Messenger Service* depicted as a messenger carrying a parcel while being

propelled by a rocket. A *noninteractive* image consisted simply of a picture of the product (or service) together with a written form of the brand (or company) name, or vice versa (e.g., *OBear Abrasive Saws* accompanying a picture of a boy holding up a large letter O).

Results clearly indicated superiority of interactive imagery in assisting recall of brand (or company) name associations with corresponding products and/or services. Lutz and Lutz (1977) noted that evidence on the superiority of visual memory over verbal memory justified pictorial advertising. However, mere use of pictorial material did not assure better recall of brand (or company) names. "An interactive image facilitates recall better than a noninteractive image, presumably by increasing the concreteness of the material to be learned, that is, the *association* of the two items. The more concrete the association becomes, the more memorable it is" (p. 497).

Facilitative effects of concrete (i.e., perceptual, action related, or emotional as distinct from conceptual) associations on brand or company name recall have far-reaching implications regarding design of brand or company name symbols and marks or ways in which advertising images are constructed. Emotional associations, in particular, are the focus of the present study and their implications regarding product preference (rather than recall of product and brand name pairings) are considered below.

The rationale for the proposed approach is that any product conveys a wide range of connotations to consumers. To enhance product appeal to consumers, a name can be selected that conveys a desirable subset of those connotations. One way to identify the desirable connotations of each new product is to conduct a separate and new survey for that product using an intuitively and arbitrarily selected list of characterizations

Correspondence concerning this article should be addressed to Albert Mehrabian, Department of Psychology, University of California, Los Angeles, 405 Hilgard Avenue, Los Angeles, California 90024.

(connotations). The latter approach would be cumbersome in addition to being selective or biased.

The proposed approach provides a systematic basis for a comprehensive characterization of basic connotations of a name. In addition, it provides measures of discrepancies between the ideal set of connotations for a product and the connotations actually implied by a given product name.

Much of the arbitrariness and difficulty in identifying connotations of a product is eliminated through the assessment of emotional connotations. Emotional connotations of a product, or any stimulus, represent the lowest common denominators of cognitive response to that product or stimulus. Furthermore, factor-analytic studies have shown that emotional connotations of stimuli (or alternatively, emotional reactions) can be characterized succinctly in terms of three dimensions: pleasure–displeasure, arousal–nonarousal, and dominance–submissiveness (note reviews provided by Mehrabian, 1980, chaps. 2 and 3). *Arousal–nonarousal* constitutes a physiological dimension characterizing the level of physical activity and mental alertness of an organism. The cognitive counterpart of arousal–nonarousal is *information rate*. Briefly, complexity, novelty, variability, and unpredictable quality of a stimulus jointly characterize its information rate and are positive correlates of the arousal elicited by the stimulus (Mehrabian, 1980, chap. 8). *Dominance–submissiveness* refers to a feeling of power, control, or influence versus a lack of power and inability to control or influence a situation.

Combinations of various levels of pleasure, arousal, and dominance are necessary and sufficient to describe any emotional state. Indeed, these three basic dimensions account for almost all of the reliable variance in existing measures of emotional states (Russell & Mehrabian, 1977).

The factors of *evaluation*, *activity*, and *potency* identified by Osgood, Suci, and Tannenbaum (1957) are analogous to the emotion dimensions of pleasure, arousal, and dominance, respectively. In attempting to identify basic dimensions of meaning, Osgood et al. helped clarify the lowest common denominators of cognitive reactions to stimuli; these common denominators were conceptualized best as connotations of, or affective and emotional associations to, the stimuli. Thus, obliquely, the semantic differential factors provided assessments of affective or emotional reactions.

Because they were designed specifically to assess emotional reactions, the verbal-report measures of pleasure, arousal, and dominance (Mehrabian, 1978) were used in our study to assess emotional connotations of products and those of product names. Future research may clarify the relative effectiveness of the semantic differential and emotion-based approaches for selecting product names.

An important refinement in the application of the three-dimensional emotion framework was assessment of the positive subset of emotional connotations for each product. Subjects used the pleasure, arousal, and dominance scales to report how they ideally would like to feel while using a product. Next, subjects used the same three scales to rate the emotions connoted by each of several names assigned to that product. It was hypothesized that *discrepancies* (on the pleasure, arousal, and dominance scales) between the ideally desired emotional impact of a product and the emotional impact of a selected name

are negative correlates of preference for the product assigned that name.

Plan of Study

The present study was conducted in three parts. Part 1 was designed to assess the emotion constellation that consumers ideally would wish to experience while using a given product. In this part, groups of male and female subjects used the three emotion scales to rate how they ideally would like to feel while using each of five representative product categories (aspirin, candy bar, car, toothpaste, wristwatch). These ratings are referred to as *desired ideal emotional impact of product* or *ideal impact*.

The five product categories used in Part 1 and throughout the study were not selected randomly; instead, they were selected carefully to represent a broad and diverse range of consumer durable and nondurable products. It was thus hoped that the diversity of products selected would enhance potential generalizability of the present findings to the universe of consumer products.

In Part 2, each of the five product categories was assigned approximately eight different names. For each product category (e.g., wristwatch), a list of product names was devised for male, and separately for female, consumers. The names for each product were selected on an a priori basis to have increasingly different emotional connotations relative to the desired ideal emotional impact of that product. Subjects in Part 2 were presented with a product category (e.g., wristwatch) together with a specific name for the product (e.g., Alert) and used the three emotion scales to rate how they would feel while using the given product of that name. These ratings are referred to as *actual emotional impact of product plus name* or *name impact*.

In Part 3, subjects of both sexes rated how much they would prefer (i.e., like and desire to purchase) a product with a given name. These *preference for product of given name* or *preference* ratings were obtained for all product and name combinations used in Part 2.

Method

Part 1: Desired Ideal Emotional Impact of a Product

Subjects. In Part 1, subjects were 50 University of California, Los Angeles (UCLA) undergraduates (25 male, 25 female) who participated in the study as part of a course requirement.

Procedure. Subjects were run in two group sessions with approximately half of the subjects in each group. Each subject received a set of five rating sheets corresponding to each of five product categories: aspirin, candy bar, car, toothpaste, wristwatch. Instructions accompanying each rating sheet are exemplified by the following for the product toothpaste.

We want you to rate below how the ideal toothpaste for you would feel while you use it. You need to take some time to imagine carefully and clearly the ideal toothpaste for yourself. Only when you have done so, will you be in a position to rate precisely how that toothpaste would make you feel.

The remaining instructions and scales on each rating sheet consisted of three basic measures of emotional state: pleasure–displeasure,

arousal–nonarousal, and dominance–submissiveness (Mehrabian, 1978; Mehrabian, 1980, chaps. 2 and 4).

Items of the three emotion-state scales are in semantic differential format. The pleasure–displeasure measure contains 24 items exemplified by the pairs of words *affectionate–nasty* and *excited–enraged*. For each pair, subjects place a check mark in one of the nine spaces separating the pair to show how they feel. The arousal–nonarousal measure contains 8 items exemplified by *troubled–dull* and *frustrated–sad*. The dominance–submissiveness measure contains 15 items exemplified by *masterful–fascinated* and *violent–fearful*. Half the items in each of the pleasure and arousal measures and 7 of the 15 dominance items are inverted to control for response bias, and items from all three scales are presented in a random order.

In the course of development of the three emotion scales, item analyses and selections in several consecutive stages resulted in the elimination of different numbers of items in each scale. Thus, only 8 satisfactory items were retained for the arousal–nonarousal scale, 15 for the dominance–submissiveness scale, and 24 for the pleasure–displeasure scale. The numbers of items in these three scales corresponded roughly to the degree of ease versus difficulty in measuring each underlying dimension. The pleasure dimension was the easiest to measure and yielded the most reliable scale among the three, whereas the arousal dimension was the most difficult to measure and yielded the least reliable scale. Reliabilities of the three scales based on data from the present study are reported later in the Results section.

The order in which subjects rated the desired ideal emotional impact of each of the five products was counterbalanced across subjects. Subjects were instructed to take a break of 3 min and to sit quietly once they had completed the rating sheet corresponding to each product. Each subject also recorded his or her sex on each rating sheet.

The data matrix obtained from male subjects in Part 1 consisted of 5 (product categories) \times 3 (emotion scale scores) \times 25 (replications), with each male subject providing one complete replication. A similar 5 \times 3 \times 25 data matrix was obtained from female subjects.

Part 2: Emotional Impact of a Product With a Given Name

Subjects. In Part 2, subjects were 300 UCLA undergraduates (150 male, 150 female) who participated in the study as part of a course requirement.

Product names. Names selected for all five product categories were terms describing emotions. The names At Ease, Vigor, Overwhelm, Relax, Daring, or Impress exemplify those assigned to toothpaste for male consumers. The names Quiet, Frenzy, Subdue, Overwhelm, and Jitters exemplify those assigned to aspirin for female consumers.

For each sex and each product, approximately eight different names were selected on an a priori basis to have increasingly different emotional connotations relative to the ideal impact rating of the product. A priori selection of names to match, or differ from, ideal impact ratings was based on available ratings of emotion terms (Russell & Mehrabian, 1977; Mehrabian, 1980, Table 3.4). Thus, for instance, two names were selected to have emotional connotations that matched as closely as possible, the ideal impact rating of the product. Two additional names had emotional connotations that differed slightly from the ideal impact rating. A third pair of names had emotional connotations that differed moderately, and a fourth pair differed strongly, from the ideal impact rating.

For the product *toothpaste*, for instance, the ideal impact ratings by both sexes had indicated moderately high pleasure, low arousal, and moderately high dominance as the desired ideal emotional impact of the product. The first pair of names designed to match the ideal impact ratings was At Ease and Relax. The second pair, differing slightly from the ideal impact ratings, was Vigor and Daring. The third pair, differing

moderately, was Overwhelm and Impress. The fourth pair, differing strongly, was Terrify and Humiliate.

Ratings of emotion terms on the pleasure, arousal, and dominance scales given by Mehrabian (1980, Table 3.4) were tantamount to general definitions of those terms and provided only approximate guidelines for our selections of product names. Even though the previous ratings of emotion terms may not have generalized exactly to product names, they nevertheless were sufficient to permit selection of several names for each product so that discrepancies between name impact and ideal impact would vary. Actual discrepancy scores for all product–name combinations were calculated specifically from ratings obtained in the present study.

Procedure. Subjects were run in groups, with 18 to 25 subjects in each group. Each subject rated half of the names assigned to a single product. For example, one subset of male subjects rated the names At Ease, Overwhelm, Relax, and Daring for the product toothpaste. Thus, each subject received a set of four rating sheets corresponding to each of four product names.

For male subjects, the product *wristwatch* had been assigned only seven names, instead of the usual eight. Thus, whereas almost all male subjects rated four different names for a single product, a few subjects rated only three different names from among those assigned to wristwatch.

As in the case of the male subjects, each female subject rated half of the names assigned to a single product. Because the product *aspirin* had been assigned only six names, instead of the usual eight, a few female subjects rated only three different names for aspirin. The remaining female subjects each rated four different names assigned to one of the remaining products.

Instructions accompanying each rating sheet are exemplified by the following for the product toothpaste named Relax.

We want you to rate below how a *toothpaste* named *Relax* would make you feel while you use it. You need to take some time to imagine carefully and clearly this toothpaste named Relax and how it would feel to use it. Only when you have done so, will you be in a position to rate precisely how this particularly toothpaste would make you feel.

The remaining instructions and scales on each rating sheet consisted of the three state measures of pleasure–displeasure, arousal–nonarousal, and dominance–submissiveness (Mehrabian, 1978; Mehrabian, 1980, chaps. 2 and 4). Evidence from earlier studies that used the emotion scales indicated that subjects could provide reliable ratings of various stimuli when given instructions similar to the preceding ones. Specific internal consistency (reliability) estimates for the three emotion scales as used with the present set of instructions are given later in the Results section.

The order in which subjects rated the product and name combinations assigned to them was counterbalanced approximately across subjects. Subjects were instructed to take a break of 3 min and to sit quietly once they had completed the rating sheet corresponding to each product and name combination. Each subject also recorded his or her sex on each rating sheet.

A total of 39 product–name combinations were rated by male subjects, with eight different product names corresponding to each of aspirin, candy bar, car, and toothpaste, and seven names for wristwatch. Each product–name combination was rated by approximately 15 male subjects.

For female subjects, a total of 38 product–name combinations were rated, with eight different product names corresponding to each of candy bar, car, toothpaste, and wristwatch, and six names for aspirin. Each product–name combination was rated by approximately 15 female subjects.

In sum, the data matrix obtained from male subjects in Part 2 consisted of 5 (product categories) \times y (product names) \times 3 (emotion scale

scores) $\times x$ (replications). The value of y was 8 for all product categories except for wristwatch where y was 7. The value of x approximated 15, with each male subject providing half the data in a replication. A similar $5 \times y \times 3 \times x$ data matrix was obtained from female subjects where y was 8 for all product categories except for aspirin where y was 6. The value of x approximated 15, with each female subject providing half of the data in a replication.

Part 3: Preference of a Product With a Given Name

Subjects. In Part 3, subjects were 293 UCLA undergraduates (145 male, 148 female) who participated in the study as part of a course requirement.

Product Preference Scale. A 12-item preference scale was used to measure consumer preferences for various products. Items of the preference scale were adapted from a measure of consumer preferences for video games reported by Mehrabian and Wixen (1986). Examples of positively worded items from the scale are "This product would offer me more than the same kind of product I currently use" or "If a friend asked me what product to buy, I would recommend this one." Responses to each item were obtained on a 9-step Likert scale. The two ends of the Likert scale were anchored with *disagree* and *agree* for both positively worded items noted. Negatively worded items are exemplified by "The product was not developed and manufactured by qualified professionals" or "This product would not satisfy my needs." Anchor words on the response scale were *was* and *was not*, and *it would not* and *it would*, respectively, for the latter two negatively worded items. Positively and negatively worded items were presented in a random order.

Procedure. Subjects were run in groups with 16 to 25 in each group. Each subject rated his or her preferences for half of the product-name combinations in a single product category (e.g., car). For example, one subset of male subjects rated their preferences for toothpastes named At Ease, Overwhelm, Relax, and Daring. Thus, each subject received a set of four preference rating sheets corresponding to each of the four product names.

For male subjects, the product *wristwatch* had been assigned only seven names, instead of the usual eight. Thus, whereas almost all male subjects rated their preferences for four different product-name combinations of a single product, a few reported preferences for only three different product-name combinations involving wristwatch.

As in the case of male subjects, each female subject also rated her preference for half of the product-name combinations in a single product category. Because the product *aspirin* had been assigned only six names, instead of the usual eight, a few female subjects provided ratings of their preferences for only three different product-name combinations involving aspirin.

Instructions accompanying each rating sheet are exemplified by the following for the product *toothpaste* named *Relax*.

We want you to rate below how much you like and would want to buy the *toothpaste* with the name *Relax*. For each question, put a check mark in one of the spaces (example: ____: ☒: ____) to show your attitude toward the toothpaste named Relax. The more you feel toward one extreme, the closer you should put your check mark to it. If you do not feel more one way than the other, put a check mark in the center, or neutral, position.

The 12 items of the preference scale followed. The order in which subjects rated their preferences for the four (or three) products assigned to them was counterbalanced approximately across subjects. Subjects were instructed to take a break of 3 min and to sit quietly once they had completed each rating sheet. Each subject also recorded his or her sex on the rating sheets.

The data matrix obtained from male subjects in Part 3 consisted of 5 (product categories) $\times y$ (product names) $\times 12$ (preference item

scores) $\times x$ (replications). The value of y was 8 for all product categories except for wristwatch where y was 7. The value of x approximated 15, with each male subject providing half the data in a replication. A similar $5 \times y \times 12 \times x$ data matrix was obtained from female subjects where y was 8 for all product categories except for aspirin where y was 6. The value of x approximated 15, with each female subject providing half the data in a replication.

Results and Discussion

Internal Consistencies and Intercorrelations of the Pleasure, Arousal, and Dominance Scales

The 300 subjects in Part 2 each rated four different product and name combinations (with a few rating only three such combinations). A total of 1,147 product and name combinations were thus rated by subjects on each of the pleasure, arousal, and dominance scales. Based on these data, the KR-20 internal consistency (reliability) coefficient (Kuder & Richardson, 1937) of the 24-item pleasure-displeasure scale was .97. The corresponding KR-20 reliability coefficient of the 8-item arousal-nonarousal scale was .81 and that of the 15-item dominance-submissiveness scale was .90.

For data obtained in Part 2, the pleasure scale correlated $-.39$ ($p < .05$) with the arousal scale, and $-.10$ ($p < .05$) with the dominance scale. The arousal and dominance scales correlated $.16$ ($p < .05$). For comparison purposes, it is useful to note that when the scales were used to rate emotional characteristics of a large and diverse sample of persons, the pleasure scale correlated $-.05$ ($p > .05$) with the arousal scale and $-.03$ ($p > .05$) with the dominance scale. The arousal scale correlated $-.06$ ($p > .05$) with the dominance scale (Mehrabian, 1978, Table 1). In a study in which subjects rated the emotional impacts of video games (Mehrabian & Wixen, 1986, p. 12), the pleasure scale correlated $-.04$ with the arousal scale ($p > .05$) and $.04$ ($p > .05$) with the dominance scale. The arousal and dominance scales correlated $.16$ ($p < .05$). Thus, intercorrelations among the three emotion scales varied depending on the sample of stimuli rated. For the particular sample of brand names used in our study, the correlation between pleasure and arousal ratings was greater than what is usually obtained. The magnitude of this correlation, however, did not present any obstacles to an adequate test of the proposed hypothesis.

Item Analysis of the Preference Measure

The preference scale which consisted of 12 items was factor analyzed and a principal component solution was obtained. There was one factor with eigenvalue exceeding 2.0 and it accounted for 74% of the total variance. This result indicated that items of the preference scale constituted an internally consistent measure of preference versus lack of preference for various consumer products.

Correlations were obtained between total preference scores and item scores across all products and subjects. These item-total correlations ranged in absolute value from .44 to .85. The item with the lowest item-total correlation, .44, was not considered satisfactory and was eliminated from the scale, resulting in a final, 11-item preference scale. The remaining item-total correlations ranged in absolute value from .64 to .85. Addi-

tional confirmation of the very high internal consistency of the 11-item preference scale was a .99 KR-20 reliability coefficient (Kuder & Richardson, 1937). Mean and standard deviation for the final preference scale were -2.7 and 24 , respectively.

Sample preference scores are given in Table 1 for a variety of product and name combinations.

Computations of Discrepancy Scores

Part 1 provided desired ideal emotional impact of product (or ideal impact) ratings. Part 2 provided actual emotional impact of product plus name (or name impact) ratings. Part 3 provided preference for product of given name (or preference) ratings. The object of the data analyses was to relate discrepancies between name impact and ideal impact ratings with preference ratings.

Computation of discrepancy scores first required standardization of the pleasure, arousal, and dominance measures. Separate sets of norms for subjects of each sex were available for the scales and were based on reports of emotional states across representative and diverse everyday situations. Although these norms were not developed in particular reference to emotional reactions to product use, they were deemed reasonably satisfactory in that most everyday situations and activities do in fact involve the use of a great variety of products.

Ideal impact ratings for each of five products were obtained from subjects of both sexes in Part 1. Male and female subjects' raw scores on the three emotion-state measures were standardized using the male and female population norms, respectively. For each product and each sex, average standardized pleasure, arousal, and dominance scores were computed next.

Name impact ratings for each of five products with approximately eight different names per product (a total of 39 product-name combinations for male subjects and a total of 38 for female subjects) were obtained in Part 2. Once again, raw scores on the three emotion-state measures were standardized separately for male and female subjects using male and female population norms, respectively. For each sex, product, and name, average standardized pleasure, arousal, and dominance scores were computed next.

Discrepancy scores for each product-name combination were computed next and consisted of absolute differences between average standardized name impact and ideal impact ratings on each dimension. For example, the pleasure discrepancy score for each product-name combination was the absolute difference between average standardized name impact and ideal impact ratings on the pleasure dimension. Sample discrepancy scores on the pleasure, arousal, and dominance dimensions are given in Table 1 for a variety of product-name combinations.

Regression Analyses

Separate multiple-regression analyses were performed on data obtained from male and female subjects to assess possible, emotion-based sex differences in product preference. In each analysis, product preference, the dependent variable, was explored as a function of discrepancy scores on pleasure, arousal, and dominance (see Table 2).

Dominance discrepancy (i.e., discrepancy between name im-

Table 1

Discrepancies (Between Name Impact and Ideal Impact) and Preference of Product

Product name	Pleasure discrepancy	Arousal discrepancy	Dominance discrepancy	Preference
Male subjects				
Car				
Endurer	2.46	0.10	0.51	1.37
Modest	1.56	1.40	1.95	0.61
Feeble	2.08	0.69	3.22	-0.96
Aspirin				
Tranquil	0.29	0.74	0.57	1.28
Arouse	1.37	1.46	0.76	-0.47
Alert	0.79	1.86	1.33	-1.05
Toothpaste				
At Ease	0.01	0.13	0.89	0.36
Daring	1.26	0.88	1.11	0.25
Overwhelm	1.89	0.21	0.72	0.13
Candy bar				
Spare Time	0.03	0.70	0.56	0.84
Excite	1.22	0.58	0.41	0.73
Amaze	1.49	0.34	0.80	0.52
Wristwatch				
Nonchalant	1.03	1.80	0.42	0.49
Startle	2.12	0.71	0.09	-0.07
Meek	1.34	0.22	2.25	-0.37
Female subjects				
Car				
Easy	0.34	0.96	0.39	1.03
Tame	1.76	0.79	2.14	0.11
Terror	3.79	0.44	0.16	-1.11
Aspirin				
Quiet	0.10	1.01	0.09	1.24
Subdue	1.68	0.94	1.27	0.80
Frenzy	1.34	2.28	0.67	-1.07
Toothpaste				
At Ease	0.27	0.23	0.64	0.52
Impress	0.74	0.62	0.65	0.67
Overwhelm	0.99	0.83	0.55	0.23
Candy bar				
Relief	0.77	0.32	0.10	0.55
Calm	0.58	1.71	0.39	0.98
Embarrass	2.87	1.11	0.60	-0.56
Wristwatch				
At Ease	0.13	0.94	0.31	0.44
Impress	0.87	0.53	0.94	0.30
Infatuate	0.88	0.20	2.26	-0.02

Note. Discrepancy scores for each product-name combination are absolute differences between average standardized name impact and ideal impact ratings on each emotion dimension. Preference scores are based on a standardized scale with mean equal to zero and standard deviation equal to one.

pact and ideal impact ratings on the dominance dimension) was a significant factor in the product preferences of males but was not significant for females. Thus, dominance discrepancy was more important in determining product preferences of male than of female subjects. On the other hand, pleasure discrepancy had a larger impact on product preferences of female than of male subjects.

Effects of discrepancy scores on product preferences of both male and female subjects were of considerable magnitude: discrepancy scores accounted for 30% of variance in product pref-

Table 2
Beta Weights Obtained From Multiple-Regression Analyses
of Male and Female Subjects' Product Preference
as a Function of Discrepancy Scores

Measure	Subjects		
	Male	Female	Both
Pleasure discrepancy	.23	.55	-.39
Arousal discrepancy	-.21	-.16	-.21
Dominance discrepancy	.35	ns	-.18
R coefficient	.55	.61	.55

Note. Each column represents the three beta weights and the multiple regression coefficient obtained from a single regression analysis. The third column represents the results of the regression analysis for the combined male-female sample. All data are computed for standardized variables to facilitate comparisons of the relative magnitudes of effects. All $ps < .05$.

ferences of male subjects and 37% of variance for female subjects. One tentative conclusion from the differing amounts of variance accounted is that female subjects were more sensitive to the emotional connotations of product names than their male counterparts.

Considering that only five categories of products were tested and subjects were college students, the preceding sex differences may not generalize to other product categories or to different consumer samples. If, however, additional research with different products and consumers corroborates the sex differences identified here, the findings could have far-reaching implications for the differential naming of products aimed at the male and female markets.

Results obtained for data of both sexes combined are given in the third column of Table 2.

The latter are reported for their possible use in situations in which the same products are designed for consumers of both sexes. Discrepancy scores accounted for 30% of variance in product preferences for the combined sample of male and female subjects.

The proposed theoretical model for fitting brand names to products received strong support from the results and provides a unique, new approach to the selection of brand names. The proposed approach and associated measures can be used in parallel fashion to select trademarks (symbols) and corporate names having the desired set of emotional connotations.

In closing, it is important to explain the selection of emotion labels as names for products in the present study. Emotion labels most directly connote specific emotions and systematic rat-

ings of a large number of such labels were available from earlier studies. The alternative would have been the use of more conventional product names (e.g., Jaguar for an automobile, Pampers for diapers). In that case, it would have been necessary first to obtain ratings of the emotional connotations of a large sample of such conventional product names so as to select the few names with desired discrepancies in emotional connotations relative to ideal impact ratings.

Practical applications of the proposed model, however, readily could use terms available in a language (e.g., Cascade, Limelight) or arbitrary or nonsense terms (e.g., Zybex). When a manufacturer or company has a set of potential names for a new product or service (or alternatively, a company has a potential set of new corporate names for itself), each of the names can be rated to assess its emotional connotations. Indeed, when a product is designed for a particular target group (e.g., blue-collar male consumers), a sample of that target group could provide the required ratings of emotional connotations. A comparable sample also could provide ratings of the emotion they would ideally wish to experience while using that product (i.e., the ideal impact rating). For each set of potential names, discrepancies between name impact and ideal impact ratings could be computed and names with the lowest discrepancy scores thus could be identified for final selection.

References

- Block, C. E. (1969). Symbolic branding: The problem of mistaken identity with design marks. *Trademark Reporter*, 59, 399-413.
- Kanungo, R. N. (1968). Brand awareness: Effects of fittingness, meaningfulness, and product utility. *Journal of Applied Psychology*, 52, 290-295.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika*, 2, 151-160.
- Lutz, K. A., & Lutz, R. J. (1977). Effects of interactive imagery on learning: Application to advertising. *Journal of Applied Psychology*, 62, 493-498.
- Mehrabian, A. (1978). Measures of individual differences in temperament. *Educational and Psychological Measurement*, 38, 1105-1117.
- Mehrabian, A. (1980). *Basic dimensions for a general psychological theory*. Cambridge, MA: Oelgeschlager, Gunn & Hain.
- Mehrabian, A., & Wixen, W. J. (1986). Preferences for individual video games as a function of their emotional effects on players. *Journal of Applied Social Psychology*, 16, 3-15.
- Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.
- Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11, 273-294.

Received June 20, 1985

Revision received January 30, 1986 ■

Effects of Negative Ions on Cognitive Performance

Robert A. Baron
Purdue University

Male and female subjects (undergraduate students) participated in two studies designed to investigate the impact of negative air ions on cognitive performance. In the first experiment, they worked on three different tasks (proofreading, memory span, word finding) in the presence of low, moderate, or high concentrations of such ions. Results indicated that among men, performance on two of these tasks (proofreading and memory span) was enhanced by moderate but not by high concentrations of ions. In the second experiment, undertaken to extend the generality of these initial results, male and female subjects performed two additional tasks (letter copying, decision making) in the presence of low, moderate, or high concentrations of ions. Output on the letter copying task increased significantly as ion level rose among both sexes. With respect to decision making, the tendency of male (but not female) participants to select initially preferred alternatives was significantly enhanced by moderate concentrations of negative ions. Together, the findings of these studies suggest that negative air ions can indeed exert appreciable effects on cognitive performance. However, contrary to claims often associated with advertising for commercially produced ion generators, these effects are neither simple nor uniformly beneficial in nature.

A growing body of evidence suggests that several aspects of the physical environment can strongly affect performance on work-related tasks. For example, such factors as heat (Bell, 1978; Fine & Kobrick, 1978), noise (Cohen, Evans, Krantz, Stokols, & Kelly, 1981), air pollution (Evans & Jacobs, 1981), and even chemicals present in drinking water (Rotton, Tikofsky, & Feldman, 1982) have all been found to exert such effects. The present experiment was designed to add to this growing body of knowledge by examining the potential impact of another environmental variable—one that has received far less attention in this regard: atmospheric electricity (as reflected in the number of positive and negative ions present in a given locale; Frey, 1961).

At first glance, the investigation of such effects might seem to be of little practical value. In fact, however, there are several grounds for seeking such knowledge. First, several industrial settings expose employees to unusually high concentrations of positive or negative air ions. For example, the air surrounding high voltage wires often becomes unusually rich in negative ions. Similarly, certain types of electronic equipment can rapidly deplete the level of negative ions in their immediate vicinity

(cf. Soyka, 1977). Because employees are exposed to such conditions repeatedly and over extended periods of time, investigation of their potential impact on task performance seems justified. Second, advertising for commercially produced ion generators often claims that these devices exert uniformly positive effects on employees' moods, alertness, and overall efficiency. Given that thousands of such generators are sold each year, it seems important to obtain scientific evidence concerning their actual efficacy.

At present, two lines of evidence offer support for the view that atmospheric electricity can influence human behavior. The first of these involves epidemiological data. In several parts of the world, winds associated with an increase in the concentration of positive air ions blow at various times during the year (e.g., the Santa Ana in California, the foehn in Europe). During such periods, industrial accidents, suicides, and even some types of crime increase in frequency (Muecher & Ungeheuer, 1961; Sulman, Pfeifer, & Hirschman, 1964). Although such findings are suggestive they are, of course, far from conclusive. The winds mentioned earlier are accompanied by other shifts in climatic factors (e.g., a slight rise in temperature, decreases in humidity). Thus, these factors, not shifts in ion concentration, may account for the effects just noted.

A second and perhaps more convincing source of evidence for the impact of ions on behavior is provided by the results of recent laboratory experiments on this topic (e.g., Baron, Russell, & Arms, 1985; Charry & Hawkinshire, 1981). In contrast to earlier investigations that often suffered from serious procedural flaws (e.g., the absence of standardized methods for generating or measuring ions), this latter group of experiments has been conducted under more carefully controlled conditions. Yet, the findings obtained in these studies indicate that atmospheric electricity can indeed influence human behavior. Specifically, exposure to such ions has been found to affect mood

This manuscript was prepared while the author was visiting professor of Management and Organization at the University of Washington.

The author wishes to express sincere thanks to Lisa Houk, Lisa Lutterbaugh, Joy Ross, and Tim Thurman for their able assistance in collection of the data, and to James Neely and Howard Weiss for their insightful comments on an earlier version of this article. Thanks are also due to David Myers for providing the choice-dilemma items, and to Gordon Russell and Robert Arms for providing the experimental apparatus.

Correspondence concerning this article should be addressed to Robert A. Baron, Department of Psychological Sciences, Purdue University, West Lafayette, Indiana 47907.

(Baron et al., 1985), physiological arousal (Charry & Hawkinshire, 1981), and even interpersonal aggression (Baron et al., 1985).

The present studies were designed to extend these previous results by examining the impact of negative ions on the performance of several cognitive tasks (cf. Wofford, 1966). In the past, such ions have generally been credited with exerting largely beneficial effects in this regard (Sulman et al., 1964). However, recent investigations (Baron et al., 1985) indicate that the impact of such particles may be quite complex in nature. Thus, no firm predictions concerning their influence on performance of the tasks used were formulated. Rather, the present investigations were largely exploratory in nature and were conducted primarily to determine whether negative ions can in fact influence cognitive performance under carefully controlled conditions.

Study 1

Method

Subjects and Design

A total of 142 undergraduates (76 women, 56 men) participated in the study. Subjects took part in the investigation in order to satisfy a course requirement.

A $3 \times 2 \times 2$ design based on three levels of negative ions (low, moderate, high), sex of subjects, and sex of experimenter was used. Subjects were randomly assigned to one of the three ion conditions, under the restriction that the proportion of men and women in each of these conditions be approximately constant.

Apparatus

The apparatus for generating negative ions was a Biotech Bionaire 1000 negative ion generator. A Biotech BT-400 Electroscop was used to measure ion levels throughout the project.

Procedure

When subjects reported to the laboratory, they were informed that the study was concerned with physiological reactions during problem-solving and related activities. They were then told that they would perform several different tasks and that as they did so, their physiological reactions would be recorded through an electrode attached to their legs. This electrode was then put in place just above the subject's ankle. In reality, it was used solely for the purpose of electrical grounding. Its presence was necessary for effective manipulation of the ion variable (cf. Charry & Hawkinshire, 1981). As the electrode was attached, the experimenter remarked in a very casual manner, that the box in the corner (the ion generator) was an air filter. The experimenter explained that there was a lot of dust in the laboratory and that it interfered with the physiological recordings. Ostensibly, the air filter helped alleviate this problem.

Time-filler tasks. At this point, subjects performed two time-filler tasks, included in the study to insure that subjects were exposed to the appropriate level of negative ions for approximately 15 min prior to the first cognitive task. The first time filler consisted of an abbreviated form of the Jenkins Activity Survey (Jenkins, Zyzanski, & Rosenman, 1979). The second time filler involved subjects' examination of 10 color pictures showing scenery and abstract art. It was explained to subjects that their physiological reactions to these stimuli would provide a baseline for further recordings. All of the pictures were neutral in content and

had been used in previous research as a baseline control (e.g., Baron & Bell, 1977).

Cognitive tasks. After the conclusion of the 15-min time-filler period, subjects performed three cognitive tasks, presented in totally counter-balanced order. One of these was a simple proofreading task. Subjects were given a typed passage and asked to circle every error that they noticed in spelling, typing, punctuation, or grammar. The passage contained many errors of each kind, and subjects worked on the task for 5 min. Dependent measures were the total number of lines they completed during this period and the percentage of the errors that they noticed in these lines.

Subjects also performed a verbal problem-solving task (the word-finding task; Reitan, 1972). In this task, they were presented with a series of sentences describing an object. In each case the name of the object was replaced by the nonsense word *grobnick*. Subjects' task was that of identifying the missing word as quickly as possible. For example, the correct solution to one item was the word *cats*. The first clue stated "Most grobnicks have long tails." The second noted that "Grob-nicks of a certain breed have very short tails." The third clue stated that "Grob-nicks can climb trees." A total of five clues were supplied for each item, and subjects were free to make as many guesses as they wished after each clue. Dependent measures of performance on this task were the mean number of guesses made by subjects per item and the mean number of clues they received prior to attaining the correct solution.

The third task was simply a measure of subjects' memory span for single digits. The experimenter began by reading lists of three digits and then proceeded, gradually and systematically, through lists of four, five, six, seven, eight, and nine digits. (There were four trials at each list length.) After each list, subjects attempted to repeat the numbers they had just heard. To be scored as correct they had to repeat every number on the list in the correct order. Memory span was defined as the number of digits subjects could recall perfectly 50% of the time.

Level and measurement of negative ions. The level of negative ions to which subjects were exposed was varied by means of the Bionaire 1000 ion generator. This equipment was placed on a table approximately 1 m from where the subject sat, with its air flow directed into their breathing space. As noted previously, the experimenter explained the presence of this device by remarking that it was an air filter, necessary for removing dust that interfered with the physiological recordings from the air. In the low (ambient) ions condition, only the fan on this apparatus was switched on. Thus, the level of negative ions present in the room remained at ambient levels ($1.0\text{--}2.0 \times 10^2/\text{cm}^3$). In the moderate ions condition, the ion-generating portion of the equipment was switched on and adjusted so that the ion level within subjects' breathing space was approximately $4.0 \times 10^4/\text{cm}^3$. Finally, in the high ions condition, the generator was adjusted so that the concentration of negative ions in subjects' breathing space was approximately $7.0\text{--}8.0 \times 10^4/\text{cm}^3$. Readings of ion concentration were taken before and after each session with a Biotech BT-400 Electroscop. These confirmed that three distinct and relatively constant levels of ions were produced throughout the project. (The mean readings in each condition, rounded to the nearest 10, were as follows: low ions $1.1 \times 10^2/\text{cm}^3$; moderate ions, $3.86 \times 10^4/\text{cm}^3$; and high ions, $7.81 \times 10^4/\text{cm}^3$. These readings were highly similar to those used in previous research; Baron et al., 1985). In all instances, ion level was set by an assistant. Thus, the experimenter was blind to ion level throughout the study. Another environmental factor, ambient temperature, known to exert significant effects on several aspects of behavior, was held constant at comfortable levels (approximately $20^\circ\text{--}22^\circ$ celsius; Baron, 1978) throughout the investigation.

Results

Proofreading Task

Separate analyses of variance (ANOVAs) were performed on the two dependent measures of proofreading: the total number

Table 1
Mean Percentage of Errors Noticed as a Function
of Sex of Subject and Ion Level

Sex of subject	Ion level		
	Low	Moderate	High
Male	80.21 _{ac}	87.67 _{bc}	75.13 _a
Female	85.08 _{bc}	86.73 _{bc}	86.54 _{bc}

Note. Means that do not share a common subscript differ significantly ($p < .05$) by Duncan multiple-range test.

of lines read by subjects and the percentage of errors present that they noticed. The analysis on total lines completed yielded only a significant main effect for sex of subject, $F(1, 108) = 5.43$, $p < .025$. This finding reflected the fact that women proofed significantly more lines ($M = 30.28$) than did men ($M = 26.88$).

In contrast, the analysis for percentage of errors noticed by subjects yielded several significant findings. First, the main effect for sex of subject, $F(1, 108) = 6.30$, $p < .02$, and the main effect for ions, $F(2, 108) = 2.97$, $p < .05$, were significant. The main effect for sex stemmed from the fact that women noticed a higher percentage of errors ($M = 86.14$) than did men ($M = 81.55$). The main effect for ions reflected the fact that subjects performed significantly better ($p < .05$) in the moderate ions condition ($M = 87.15$) than in either the low ions ($M = 82.93$) or high ions ($M = 82.19$) conditions.

These main effects were qualified, however, by a significant interaction between sex of subject and ions, $F(2, 108) = 4.30$, $p < .02$. The means involved in this interaction are presented in Table 1. As can be seen from this table, the interaction stemmed from the fact that ion concentration affected the performance of men, but exerted no appreciable effects on the performance of women. Specifically, men performed better in the moderate ions than in the low ions condition ($p < .07$), but worse in the high ions than in the moderate ions condition ($p < .05$). In contrast, the performance of women did not differ significantly across the three ion levels.

No effects of sex of experimenter and no interactions of this factor with other variables were observed for either dependent measure.

Word-Finding Task

An ANOVA performed on the mean number of guesses made by subjects per item yielded a significant main effect only for sex of subject $F(1, 108) = 6.26$, $p < .01$. This reflected the fact that men made more guesses ($M = 2.26$) than did women ($M = 2.01$). A similar analysis on the number of clues required by subjects prior to solution of each item yielded an identical pattern of findings. That is, only the main effect for sex of subject was significant, $F(1, 108) = 4.64$, $p < .04$. This finding reflected the fact that men required slightly fewer clues ($M = 2.18$) than did females ($M = 2.44$). It should also be noted that the small number of clues required by subjects in order to attain correct solutions to each item suggests the presence of a restrictive ceil-

ing effect. Given these conditions, the absence of additional significant findings with respect to this task is not surprising.

Memory Span

An ANOVA conducted on the data for memory span yielded no significant effects. However, the interaction between sex of subject and ions closely approached significance, $F(2, 108) = 2.82$, $p < .06$. Given the exploratory nature of this initial study, it was deemed appropriate to examine this interaction more closely, despite the fact that it did not attain conventional levels of statistical significance.

The means involved in the interaction between ion level and sex of subject are shown in Table 2. As can be seen from inspection of this table, the pattern of findings closely resembled that obtained with the proofreading task. Men performed slightly (but not significantly) better in the moderate ions than low ions condition. However, they performed significantly worse in the high ions than moderate ions condition ($p < .05$). In contrast, performance by women was not significantly affected by ion level. Again, no significant effects for sex of experimenter or interactions of this variable with other factors in the study, were obtained.

Discussion

The results of this initial study offer support for the view that negative ions can, in fact, influence performance on cognitive tasks. Among men, such effects were observed with respect to proofreading, and a trend in the same direction occurred for memory span. Contrary to many claims concerning the impact of such ions, however, these shifts in performance were *not* uniformly beneficial. On the contrary, although moderate concentrations of negative ions facilitated performance on a proofreading task, high levels of such particles failed to produce such effects. In fact, they reduced performance to a level significantly below that observed in the moderate ions condition. Similarly, although moderate levels of negative ions produced modest increments in memory span, higher concentrations of such particles again failed to yield similar benefits. Rather, they reduced performance significantly, relative to that observed in the moderate ions condition. In sum, moderate levels of negative ions enhanced task performance, whereas a further increase to high ion concentration reduced performance (although not significantly) below baseline (low ion) levels and significantly below that in the moderate ions condition. This pattern of outcomes

Table 2
Memory Span as a Function of Sex of Subject and Ion Level

Sex of subject	Ion level		
	Low	Moderate	High
Male	7.15 _{ac}	7.37 _{bc}	6.54 _a
Female	7.16 _{ac}	6.94 _{ac}	7.19 _{ac}

Note. Means that do not share a subscript differ significantly ($p < .05$) by Duncan multiple-range test.

Table 3
Mean Number of Letters Copied as a Function of Sex of Subject and Ion Level

Sex of subject	Ion level		
	Low	Moderate	High
Male	27.69 _a	33.75 _a	40.50 _b
Female	27.67 _a	35.50 _b	36.92 _b

Note. For each sex, means that do not share a common subscript differ significantly ($p < .05$) by Duncan multiple-range test.

is markedly different from that often described in mass media reports dealing with the impact of such ions.

Although the findings of this initial study are suggestive, they are certainly far from conclusive. Thus, a second investigation was performed to both extend and clarify these preliminary results. Specifically, this follow-up study was designed to accomplish three goals. First, it sought to determine whether the findings obtained in the first study could be replicated with additional tasks. For this purpose, one task more complex and one task simpler than those used in Study 1 were selected. Second, it sought to determine whether the sex difference uncovered in the initial investigation (the finding that men, but not women, were affected by ion level) would appear again with these different tasks. Finally, because the findings of previous studies suggest that negative ions may exert their influence through increments in subjects' level of arousal, this follow-up experiment included one physiological measure of arousal, resting heart rate (cf. Baron et al., 1985).

Study 2

Method

Subjects, Design, and Apparatus

A total of 72 undergraduates (36 men, 36 women) participated in the study. They participated in order to satisfy a course requirement. A randomized groups design based on three levels of ions (low, moderate, high) was used. Data from male and female subjects were gathered by different investigators. For this reason, data from the two sexes were treated as replications and analyzed separately. Apparatus was the same as that used in Study 1.

Procedure

Procedures were identical to those of Study 1, with the following exceptions. First, resting heart rate was assessed at three points: on subjects' entry into the laboratory, after completion of the two time-filler tasks (10 min later), and again after completion of the cognitive tasks described below.

Second, subjects performed two different tasks that were not used in Study 1. One of these involved decision making, and used choice dilemma items of the type included in many previous studies of both individual and group decision making (e.g., Lamm & Myers, 1978). Each of these items described a hypothetical situation in which an individual faced the task of choosing between a relatively risky but attractive course of action and a conservative but less attractive one. (Example: A

man must decide whether to invest funds in a small company whose stock might rise sharply, or to leave this money in a safe but low-yielding insurance policy.) Subjects' task for each item was that of indicating the minimum probability of success they would require before recommending adoption of the risky alternative. Eight choice dilemma items were used, and these were carefully selected on the basis of previous research to be ones for which subjects' demonstrated clear initial preferences for either risk or caution. Thus, four of the items were ones for which most persons indicated an initial preference for risk (they would require a low probability of success before recommending this alternative), whereas four were items for which most persons indicated an initial preference for caution (they would require a high probability of success before recommending the riskier alternative). For each set of items, the dependent measure was the mean probability of success required by subjects. (The higher the value, the stronger was their preference for caution in their decisions.)

Subjects also performed a simple perceptual-motor task. This task required them to copy individual letters printed on a sheet of paper both upside down and backward in the spaces immediately below them. Subjects worked on this task for 3 min. Dependent measures were the number of letters they copied during this period and the number of errors made. As in Study 1, the two tasks used were presented in counterbalanced order.

Results

Letter Copying Task

Separate ANOVAS were performed on the number of letters copied by male and female subjects. Both of these analyses yielded significant main effects for ion condition, $F(2, 27) = 3.83, 8.88, p < .034, p < .001$, respectively, for women and men. The means associated with these effects are presented in Table 3. As can be seen from this table, the number of letters copied by members of both sexes increased as ion level rose.

Corresponding ANOVAS were performed on the number of errors made by men and women in the copying task. These analyses yielded a significant main effect for ions only in the case of men, $F(2, 27) = 3.62, p < .05$. As shown in Table 4, men made fewest errors in the low ions condition ($M = 2.75$), more errors in the moderate ions condition ($M = 3.42$), and the greatest number of errors in the high ions condition ($M = 4.92$). Corresponding results were not obtained for women.

Decision-Making Task

That participants in the present study did in fact possess contrasting preferences with respect to the two sets of choice-di-

Table 4
Mean Number of Errors as a Function of Sex of Subject and Ion Level

Sex of subject	Ion level		
	Low	Moderate	High
Male	2.75 _a	3.42 _a	4.92 _b
Female	3.75 _a	2.92 _a	3.75 _a

Note. For each sex, means that do not share a common subscript differ significantly ($p < .05$) by Duncan multiple-range test.

Table 5
*Mean Preference for Caution as a Function
 of Sex of Subject and Ion Level*

Sex of subject	Ion level		
	Low	Moderate	High
Male	5.75 _{ac}	8.66 _b	7.69 _{bc}
Female	7.85 _a	7.50 _a	7.40 _a

Note. For each sex, means that do not share a common subscript differ significantly ($p < .05$) by Duncan multiple-range test.

lemma items is indicated by the fact that they reported sharply different overall levels of risk or caution for both. Among women, subjects' mean recommendations for the cautious and risky items were 7.58 and 4.24, respectively ($p < .01$). Among men, the corresponding values were 6.96 and 3.97, respectively ($p < .01$). Thus, as suggested by previous research (cf. Lamm & Myers, 1978), subjects preferred contrasting strategies for the two sets of items.

To determine whether ion level affected subjects' decisions with respect to the choice-dilemma items, separate ANOVAS (for men and women) were performed on the data for the cautious and risky sets. No significant effects were obtained for women. However, among men, the effect of ions attained significance for the caution-preferred items, $F(2, 27) = 3.80$, $p < .05$. The means associated with this effect are shown in Table 5. As can be seen from this table, men reported the strongest preference for caution in the moderate ions condition. Thus, findings for this measure were similar to those obtained in Study 1 for both proofreading and memory span.

Resting Heart Rate

Ion concentration failed to exert any appreciable effects on resting heart rate during any stage of the experiment. Thus, there was no evidence from this measure that negative ions affected subjects' level of arousal.

Discussion

The results of Study 2 both confirm and extend the findings of Study 1. As in the initial experiment, negative ions exerted significant effects on subjects' performance. Women copied a greater number of letters in both the moderate and high ions conditions than in the low ions condition, whereas men copied more letters in the high ions condition than in either the low or moderate ions conditions. Similarly, men made more errors in the high ions than in the moderate or low ions conditions.¹ Finally, male subjects' initial preferences for cautious choices in a decision-making task were enhanced by moderate concentrations of negative ions. Because the tasks used in Study 2 were specifically chosen to differ from those used in Study 1 in terms of complexity, these findings extend the initial results in both directions along this dimension.

Although the findings of Study 2 were generally consistent with those of the initial investigation, they also differed in cer-

tain respects. First, significant effects were obtained for women as well as for men on the letter copying task. This suggests that at least with respect to such simple tasks, negative ions can appreciably affect the cognitive performance of both sexes. Second, the relationship between ion concentration and task performance appeared to be linear rather than curvilinear with respect to the letter copying task. Output on this task rose for both sexes as ion level increased; there was no indication of a leveling off or actual down-turn at high ion concentrations, as was observed in Study 1. It appears, therefore, that for relatively simple tasks, increasing concentrations of negative ions may indeed exert beneficial effects on performance. (It should be noted, however, that consistent with the findings of Study 1, there was some suggestion of a curvilinear relationship between ion level and performance on the decision-making task. As may be recalled, male subjects indicated a stronger preference for initially preferred cautious choices in the moderate than high ions condition.)

General Discussion

Together, the findings of these two studies offer support for the suggestion that negative ions can affect performance on several different cognitive tasks. In Study 1, ion level affected performance on a proofreading task as well as memory span for single digits. In Study 2, this factor significantly influenced performance on a letter copying task, and on at least one aspect of decision making. Thus, it appears that negative ions can affect performance on tasks involving several different aspects of cognitive functioning.

It should be noted, however, that with the exception of the letter copying task, the above effects were generally stronger for men than for women, a finding consistent with the results of several previous studies (e.g., Charry & Hawkinshire, 1981). Full comprehension of the mechanisms responsible for this apparent sex difference are beyond the scope of this exploratory research. However, one possible explanation may be briefly mentioned.

A substantial body of literature suggests that women are better buffered physiologically than are men, and so can adapt to a wide range of environmental stressors more effectively (cf. Hoyenga & Hoyenga, 1979). Because high levels of air ions can sometimes induce pronounced negative shifts in mood (Sulman et al., 1964) and may act as a form of stress under some conditions (Muecher & Ungeheuer, 1961), it is not surprising that women were affected to a smaller degree than were men by this environmental factor in the present research. This reasoning further suggests that women, too, may be affected by negative ions, but that such effects will often be visible only at somewhat higher concentrations of such particles, or longer exposure to them, than is the case for men. This possibility can be readily

¹ That the rise in number of errors shown by men is not entirely attributable to the fact that they copied more letters as ion concentration rose is suggested by the fact that the percentage, not merely the number, of errors increased across the three experimental conditions. (The percentage of errors in the low, moderate, and high ions conditions were 9.93, 10.13, and 12.15, respectively.)

examined in further research through the use of higher ion levels and longer periods of exposure than those used here.

Another aspect of the present findings deserving of some attention is the possibility, suggested by several dependent measures, that the relationship between ion level and task performance is curvilinear in nature. Such a pattern was observed among men, for the proofreading, memory-span, and decision-making tasks. However, it failed to appear for either men or women with respect to letter copying. Given the exploratory nature of the present research, these findings should be interpreted with caution. However, the fact that the relationship between ion level and performance appeared to be curvilinear for relatively complex tasks but linear for a simpler one suggests a mechanism that may underlie the impact of negative ions on cognitive performance.

Briefly, it seems possible that rising concentrations of such particles generate increments in arousal. Because task performance often increases with arousal up to a point but then decreases as arousal continues to rise (Berlyne, 1967), this proposed mechanism would account for the curvilinear relation previously described. Further, because the inflection point on the arousal-performance function depends, in part, on task complexity, an interpretation in terms of arousal would also explain the absence of a down-turn for the letter copying task. Because this task is quite simple in scope, reductions in performance would be expected to occur only at high levels of arousal—perhaps ones above those generated in the present research.

Although an interpretation based on arousal appears to be consistent with several of the findings of the present research, it is not supported by the data for resting heart rate. Ion concentration failed to exert any significant effects on this measure. Of course, it is possible that heart rate was simply insensitive to ion-generated increments in arousal in the present context, and that some other measure of arousal would have yielded positive results. In the absence of direct evidence on this issue, however, an interpretation based upon the potential arousing properties of negative ions must be viewed as only speculative in nature. Further research is necessary before the mechanism responsible for the effects on task performance observed in the present research can be identified.

Regardless of the nature of this underlying mechanism, however, the findings of these exploratory studies have important practical implications. These appear to fall under two major categories. First, as noted earlier, devices for the generation of negative ions are commercially manufactured and actively marketed by several companies. Advertising for such devices often asserts that negative ions exert uniformly beneficial effects on persons exposed to them (e.g., increments in mental alertness and efficiency). The present findings cast serious doubt on such claims. Although high levels of negative ions enhanced performance on a simple copying task, they failed to improve accuracy in proofreading and memory span for single digits; in these latter tasks, the higher concentration of negative ions produced performance that was slightly poorer than that in the baseline (low ions) condition. Further, such environmental conditions strengthened the tendency of male subjects to select initially preferred alternatives in a decision-making task—a shift that

can be viewed as inimical to flexibility. Given that large numbers of ion generators are sold each year, these findings are of considerable interest.

Second, the present findings underscore the importance of measuring, and perhaps adjusting, ion concentrations in certain industrial settings. Previous research (cf. Soyka, 1977) suggests that several industrial processes, some types of electronic equipment, high voltage power lines, and even modern heating and air conditioning systems can all affect the number and relative concentration of ions present in a given locale. As a result, many employees are exposed to ion levels at work that differ sharply from those existing outdoors, at home, or elsewhere. The present study, and several previous experiments, indicate that short-term exposure to such conditions can produce measurable shifts in mood, task performance, and various forms of social behavior (cf. Baron et al., 1985; Charry & Hawkinshire, 1981). Given this fact, it seems possible that more prolonged exposure to these circumstances may yield effects of considerable practical importance. Systematic investigation of this possibility seems warranted.

To conclude: The present research serves to extend previous findings by indicating that atmospheric electricity (in the form of negative air ions) can affect performance on several cognitive tasks, as well as current moods (Baron et al., 1985), physiological reactions (Charry & Hawkinshire, 1981), and various aspects of social behavior (Baron et al., 1985). In contrast to statements in the popular press and in advertisements for commercially produced ion generators, however, such effects appear to be quite complex in scope, and far from uniformly beneficial in nature.

References

- Baron, R. A. (1978). Aggression and heat: The "long hot summer" revisited. In A. Baum, S. Valins, & J. E. Singer (Eds.), *Advances in environmental research* (pp. 57–82). Hillsdale, NJ: Erlbaum.
- Baron, R. A., & Bell, P. A. (1977). Sexual arousal and aggression by males: Effects of type of erotic stimuli and prior provocation. *Journal of Personality and Social Psychology*, 35, 79–87.
- Baron, R. A., Russell, G. W., & Arms, R. L. (1985). Negative ions and behavior: Impact on mood, memory, and aggression among Type A and Type B persons. *Journal of Personality and Social Psychology*, 48, 746–754.
- Bell, P. A. (1978). Effects of noise and heat stress on primary and subsidiary task performance. *Human Factors*, 20, 749–752.
- Berlyne, D. E. (1967). Arousal and reinforcement. In D. Levine (Ed.), *Nebraska Symposium on Motivation* (Vol. 15, pp. 279–286). Lincoln: University of Nebraska Press.
- Charry, J. M., & Hawkinshire, F. B. W., V. (1981). Effects of atmospheric electricity on some substrates of disordered social behavior. *Journal of Personality and Social Psychology*, 41, 185–197.
- Cohen, S., Evans, G. W., Krantz, D. S., Stokols, D., & Kelly, S. (1981). Aircraft noise and children: Longitudinal and cross-sectional evidence on adaptation to noise and the effectiveness of noise abatement. *Journal of Personality and Social Psychology*, 40, 331–345.
- Evans, G. W., & Jacobs, S. V. (1981). Air pollution and human behavior. *Journal of Social Issues*, 37, 95–125.
- Fine, B. J., & Kobrick, J. L. (1978). Effects of altitude and heat on complex cognitive tasks. *Human Factors*, 20, 115–122.
- Frey, A. H. (1961). Human behavior and atmospheric ions. *Psychological Review*, 68, 225–228.

- Hoyenga, K. B., & Hoyenga, K. T. (1979). *The question of sex differences: Psychological, cultural, and biological issues*. Boston: Little, Brown.
- Jenkins, C. D., Zyzanski, S. J., & Rosenman, R. H. (1979). *Jenkins Activity Survey*. New York: Psychological Corporation.
- Lamm, H., & Myers, D. G. (1978). Group-induced polarization of attitudes and behavior. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 145–195). New York: Academic Press.
- Muecher, H., & Ungeheuer, H. (1961). Meteorological influence on reaction time, flicker-fusion frequency, job accidents, and medical treatment. *Perceptual and Motor Skills*, 12, 163–168.
- Reitan, R. M. (1972). Verbal problem solving as related to cerebral damage. *Perceptual and Motor Skills*, 34, 515–524.
- Rotton, J., Tikofsky, R. S., & Feldman, H. T. (1982). Behavioral effects of chemicals in drinking water. *Journal of Applied Psychology*, 67, 230–238.
- Soyka, F. (1977). *The ion effect*. New York: Bantam Books.
- Sulman, F. G., Levy, D., Levy, A., Pfeifer, Y., Superstein, E., & Tal, E. (1964). Ionometry of hot, dry desert winds (sharav) and application of ionizing treatment to weather-sensitive patients. *International Journal of Biometeorology*, 18, 393.
- Sulman, F. G., Pfeifer, Y., & Hirschman, M. (1964). Effect of hot dry desert winds (sharav, hamsin) on the metabolism of hormones and minerals. *Harokeach Haivri*, 10, 401–404.
- Wofford, J. C. (1966). Negative ionization: An investigation of behavioral effects. *Journal of Experimental Psychology*, 71, 608–611.

Received January 8, 1986

Revision received February 25, 1986 ■

Instructions to Authors

Articles submitted for publication in the *Journal of Applied Psychology* are evaluated according to the following criteria: (a) significance of contribution, (b) technical adequacy, (c) appropriateness for the journal, and (d) clarity of presentation. In addition, articles must be clearly written in concise and unambiguous language. They must be logically organized, progressing from ■ statement of problem or purpose, through analysis of evidence, to conclusions and implications.

Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (3rd ed.). Articles not prepared according to the guidelines of the *Manual* will not be reviewed. All manuscripts must include an abstract of 100–150 words typed on a separate sheet of paper. Typing instructions (all copy must be double-spaced) and instructions on preparing tables, figures, references, metrics, and abstracts appear in the *Manual*. Also, all manuscripts are subject to editing for sexist language.

Authors can refer to recent issues of the journal for approximate length of regular articles. (Three double-spaced manuscript pages equal one printed page.) A few longer articles of special significance are occasionally published as monographs. Short Notes feature brief reports on studies such as those involving some methodological contribution or important replication.

APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more journals. APA policy also prohibits duplicate publication, that is, publication of a manuscript that has already been published in whole or in substantial part in another journal. Also, authors of manuscripts submitted to APA journals are expected to have available their raw data throughout the editorial review process and for at least 5 years after the date of publication.

Authors will be required to state in their initial submission letter or sign a statement that they have complied with APA ethical standards in the treatment of their sample, human or animal. A copy of the APA Ethical Principles may be obtained from the APA Ethics Office, 1200 17th Street, N.W., Washington, DC 20036.

Anonymous reviews are optional, and authors who wish anonymous reviews must specifically request them when submitting their manuscripts. Each copy of a manuscript to be anonymously reviewed should include a separate title page with authors' names and affiliations, and these should not appear anywhere else on the manuscript. Footnotes that identify the authors should be typed on a separate page. Authors should make every effort to see that the manuscript itself contains no clues to their identities.

Manuscripts should be submitted in quadruplicate and all the copies should be clear, readable, and on paper of good quality. A dot matrix or unusual typeface is acceptable only if it is clear and legible. Authors should keep a copy of the manuscript to guard against loss. Mail manuscripts to the Editor, Robert Guion, Department of Psychology, Bowling Green State University, Bowling Green, Ohio 43403.

Receptivity and Planned Change: Community Attitudes and Deinstitutionalization

Gregory H. Wilmoth, Starr Silver, and Lawrence J. Severy
University of Florida

An expectancy-value model was used to measure and explain receptivity attitudes (i.e., change climate) toward the implementation of deinstitutionalization programs. Questionnaires measuring values, expectancies, and behavioral intentions were mailed to community leaders and to members of community groups believed to be important in setting opinions and making decisions. Responses from 599 persons revealed that (a) the size of a proposed group home affected neither attitudes nor intentions of support, (b) group homes for mental health clients were viewed with less favorable attitudes and intentions than those for the retarded or the elderly, (c) members of various community groups held significantly different attitudes and intentions toward the programs, and (d) attitudes and intentions toward deinstitutionalization were more favorable than toward institutionalization. The application of this approach for assessing the implementation climate for planned change was discussed.

Successful change is more likely when the change agent has knowledge of the change climate. Who supports or opposes this change and why? How strong and widespread is the support and opposition? Is interest group membership a better predictor of receptivity than demographic variables? These are the types of questions we attempted to answer for the implementation of deinstitutionalization programs in Florida. Our purpose focused only on the nonclinical issue of receptivity to this change. We did not attempt either to change attitudes or to measure attitude change.

Successful implementation must often rely on acceptance (attitudes) and commitment (actions) from relevant interest groups and leaders both in and outside the formal organization implementing the change (Anderson & Bell, 1978; Rogers & Shoemaker, 1971). Previously demonstrated effectiveness of a planned change alone is not sufficient for its successful implementation elsewhere (Rogers & Shoemaker, 1971; Walton, 1975). Implementation includes social, psychological, and political processes as well as the technical aspect of demonstrated effectiveness (Stone, 1980).

Deinstitutionalization is an example of a largely effective (Braun et al., 1981; Kiesler, 1982) and cost-efficient policy (Murphy & Datel, 1976; Weisbrod, Test, & Stein, 1980). In spite of their demonstrated effectiveness, deinstitutionalization

efforts, especially group residential facilities, have often met with vehement, vociferous opposition (CBS, 1980). A national survey (Baron & Piasecki, 1981) found that, for every facility open at the time of the survey, another facility was either never opened or closed because of community opposition. Opposition from a few influential community leaders or an active interest group can alone be sufficient to block or evict a residential facility (Locker, Rao, & Weddell, 1979; Wilgus & Epstein, 1978; see Walton, 1975, for similar examples in organizational development). Public opposition resulted in the statewide suspension of California's deinstitutionalization program (Rabkin, 1975). Community resistance can also sabotage program success by eroding conditions that contribute to client rehabilitation (Segal, Baumohl, & Moyles, 1980; Test, 1981).

Controversy over a program arises when groups hold different attitudes toward it. Attitude differences result from differences in values and/or differences in beliefs about the consequences of the attitude object on fulfillment of these values (Fishbein & Ajzen, 1975; Rosenberg, 1956). At the outset of our study, attitudes toward the planned change (i.e., deinstitutionalization programs) represented receptivity. We defined resistance as a negative attitude toward planned change, whereas acceptance was a positive attitude. The person's attitude (i.e., evaluative response) was based on the values that the individual expected would be fulfilled by the attitude object. The sum of these expectancy-value products measured an individual's attitude. This approach allowed respondents to balance the expected benefits of change against the expected costs (Rothbart, 1973). Attitudes sometimes are accompanied by behavioral intentions to act.

Although attitudes are held by individuals, conflicts over planned change are most often intergroup disagreements. We combined an attitude model of individual receptivity with an interest group model of community decision making (O'Brien, 1975).

This research was funded by the Service Through Applied Research program of the Board of Regents, State University System, State of Florida (STAR-79-034), in cooperation with the Department of Health and Rehabilitative Services, State of Florida. Points of view or opinions in this manuscript are those of the authors and do not necessarily represent the official positions or policies of the State of Florida Board of Regents or the Department of Health and Rehabilitative Services.

Correspondence concerning this article should be addressed to Gregory H. Wilmoth, who is now at the University of Maryland University College, European Division, APO New York, New York 09102.

Acting out of self-interest, individuals resist change that they perceive as having negative consequences for themselves (Rothbart, 1973). People with shared values organize into interest groups to increase their bargaining power within the community (O'Brien, 1975). Communities and organizations consist of individuals who are members of groups with different vested interests.

Studies of attitudes toward community-based mental health programs (Miller, 1981; Shadish, Thomas, & Bootzin, 1982) showed that even proponents of these programs had different mean values. Proponents, such as clinicians, families of clients, and administrators, had different concerns (i.e., values) reflecting their divergent roles vis à vis these programs.

People with similar values, however, can possess different expectations about how their values will be affected by a program or planned change (attitude object; Mazmanian & Sabatier, 1980). For yet-to-be-implemented programs, attitudes were a function of the attributes and outcomes that the planned change was expected to possess (Severy, Houlden, & Wilmoth, 1981; Sundstrom, Lounsbury, Schuller, Fowler, & Mattingly, 1977). Some residents expected (i.e., feared) their property values, for example, to decline as a result of a proposed group home (Dean, 1977; Maypole, 1981). These fears can be a formidable psychological barrier to the implementation of planned change.

More specifically, we expected receptivity to a group home to be mediated by attitudes toward the intended clients. People attach more stigma to mentally ill patients, evaluate them more negatively, and perceive them as more unpredictable and dangerous than either elderly or retarded patients (Brockman, D'Arcy, & Edmonds, 1979; Fracchia, Canale, Cambria, Ruest, & Sheppard, 1976; Tringo, 1970). We further predicted that fears would increase as the number of clients residing in a nearby group home increased (Baron & Piasecki, 1981; Taylor, Dear, & Hall, 1979).

Although demographic variables influenced attitudes toward mental illness (Rabkin, 1972, 1975), demographic variables did not show a clear relation to attitudes toward group homes for various clients (Johnson & Beditz, 1981; Locker et al., 1979; Segal et al., 1980). For example, within the upper socioeconomic stratum, lawyers and physicians expressed different attitudes (Morrison, Smith, Fentiman, Madrazo-Peterson, & Boyagian, 1979). Therefore, we tested an interest group model of community receptivity rather than testing solely demographic variables. Demographic variables, however, were included as covariates.

To summarize, the purposes were (a) to assess the level of pre-implementation receptivity, (b) to identify the values and expectancies associated with receptivity, (c) to identify the sources of support or opposition (the relevant interest groups), (d) to determine whether selected objective attributes of the planned change influenced receptivity, and (e) to discover to what extent, if any, the planned change possessed a perceived relative advantage over the status quo. People opposed to group homes may prefer them to institutionalization (i.e., the status quo).

Our detailed hypotheses were as follow.

1. Attitudes and intentions will be more favorable for smaller than for larger group homes (i.e., resident population size).

2. Attitudes and intentions will be more favorable toward homes serving elderly clients than for those serving mental health or retarded clients.

3. Community attitudes and behavioral intentions will not be homogeneous but will be different across interest groups.

4. Attitudes and intentions will be more favorable toward group homes than toward state institutions.

5. Attitudes and intentions will be more favorable toward group homes and outpatient deinstitutionalization programs as a policy option than toward institutionalization as an option.

6. Attitudes and intentions will be affected by respondents' sex, age, education and socioeconomic status (SES).

7. Attitudes and intentions will be different across interest groups with sex, age, education, and SES partialled out.

8. Attitudes will be positively related to intentions.

Method

Respondents indicated the degree to which they valued various program attributes (i.e., values) and the extent to which they believed these values were either characteristic of or influenced by each of the programs. Values (e.g., having property values in a neighborhood decrease) were measured on a scale from *very much like* (3) to *very much dislike* (−3). Expectancies for each program (on each value) were assessed on a scale from *very likely to possess this characteristic* (3) to *very unlikely to possess this characteristic* (−3). To moderate potential contamination of the value ratings by expectancies for specific programs, values were rated separately and before expectancies. An expectancy-value score of 9 (maximum score) occurred whenever a program was seen as very likely either to facilitate (3) the attainment of a highly desirable (3) outcome, or to block (−3) the occurrence of a very undesirable (−3) outcome. Similarly, a score of −9 (minimum score) occurred whenever a program was believed either to lead to very undesirable outcomes or to block the attainment of very desirable ones. Because attitudes were expressed as the sum of the expectancy-value products, positively valenced products contributed to favorable attitudes toward a program, and negatively valenced products contributed to unfavorable attitudes. Finally, respondents expressed their intention to support or oppose each type of program if it was proposed to be built within 1 mile of their homes. The behavioral intention scale ranged from *strongly oppose* (work to block, −3) to *strongly support* (work to support, 3). All scales were 7-point, bipolar scales.

The criterion of being within 1 mile of the respondent's home was selected as a reasonable choice for use in rural areas (66% of our districts), where group homes were often isolated from nearby houses. It was a less than satisfactory choice for urban areas.

Questionnaire Development

We used extensive interviews to identify the most relevant characteristics and potential outcomes of the programs. State human service program planners identified group home programs that had been especially supported or opposed. Individuals familiar with these programs were interviewed, in both one-on-one and focus group interviews (Calder, 1977), regarding important program characteristics that should be incorporated and avoided in the operation of these programs. The focus group interviews consisted of an interviewer focusing a group's discussion on the topic of community-based mental health services. Interviewees ($N = 53$) represented the decision-making, delivery, and response systems (i.e., clients' families) involved in the operation of these programs and thus provided information about perceived relevant characteristics from diverse frames of reference.

Pilot Testing

We extracted 78 characteristics (values) likely to be associated with supported and/or opposed programs. Because many items overlapped conceptually and other characteristics seemed to be important to only a few individuals, we undertook subsequent instrument refinement. Ratings of the perceived value (but not expectancies) of the 78 characteristics were obtained from group home vendors, state human service personnel, and individuals from the private sector who attended a legislative community forum on group homes ($N = 200$). Factor analysis of these responses (principal components with varimax rotation) helped us identify and eliminate redundant items. We also discarded characteristics (items) which received average value ratings below 1.0 because they would contribute little mathematically to the computation of expectancy-value products. A few characteristics with low average value ratings were judiciously retained if we thought they had theoretical import or were likely to be of special importance to one of the targeted interest groups. Forty program characteristics were chosen for the final survey (which can be obtained from the authors).

Two aspects of the planned evaluation necessitated construction of four survey instruments. First, rating five program alternatives on 40 characteristics was too lengthy for the average individual to complete with proper attention. Hence, program characteristics were divided into two sets by matching the characteristics on their mean value rating (determined by the pilot testing). Items from each pair were randomly assigned to Value Set A or B. Items were not matched for expectancy ratings. Second, to test the influence of the size of the group home on receptivity, we constructed two descriptions of group homes. In one description, the group homes provided residential care for from 3 to 12 clients; whereas from 18 to 30 clients were served by group homes in the second description.

To determine the effect of client type and facility size on attitudes toward group homes, the descriptions of all other characteristics of the homes were held constant. The order of programs in the questionnaires was determined randomly and was constant throughout all questionnaires (i.e., group homes for the retarded, mentally ill, and elderly; state institutions, and outpatient programs). Each group home was described as a 24-hr facility that provided personal services (food, bathing, etc.) to clients who may have been previously institutionalized and were currently unable to live independently. Mentally retarded clients were described as being 18 years of age or younger, with below-normal mental abilities. On the basis of state funding categories, the mentally ill were described as having either emotional and/or drug and alcohol problems. The elderly were described as clients over 60 years of age having physical, emotional, or mental problems. Outpatient care programs were described as serving those mentally ill, retarded, and elderly clients who lived with their families, in foster homes, or otherwise not in a group home or institution. Clients in outpatient care programs were described as being unable to live independently without supervision. State institutions were described as housing several hundred mentally ill, retarded, and elderly clients. (Because of a typographical error on the surveys, small group homes for the mentally retarded were described as housing mental health clients. Thus, all data for small group homes for the mentally retarded were excluded from all analyses.)

Survey Sample

State social service planners were interested in surveying the attitudes and behavioral intentions of community members residing in three administrative districts of the state. These districts were avoided during instrument development. The final questionnaires were distributed by mail to (a) all state legislators, all mayors within each target district, and random samples of city and county commissioners (referred to in this study as government leaders); (b) state social service staffs; (c) vendored

service providers (i.e., group home operators); (d) physicians; (e) realtors; and (f) family members of clients. Surveys were also mailed to a random sample of the general public in the target districts. The random sample was not included in the analyses because our goal in this study was not to describe the attitudes of the public at large but rather to test a model of receptivity based upon interest groups. By definition (O'Brien, 1975), the public at large is not an interest group. Each individual received only one of the four questionnaires (Value Set A or B by small vs. large group home description). The instruments were stratified by type and distributed on a random basis. A reminder notice was mailed 2 weeks later.

There was variation in the return rates. Physicians were least likely (10%) and state social service staff most likely (46%) to return the instrument. The overall response rate was 21%. A total of 599 usable responses were received.

Analytic Sample

The disproportionate response rates and group sizes held the potential to skew the results. To compensate for this, a sample of returned surveys was constructed. Two criteria guided the sampling procedure. First, group size was targeted to the group with the fewest responses (physicians, $n = 50$). Second, for all groups with more than 50 respondents, only those questionnaires with no missing data were used. This resulted in one group with only 46 complete surveys. All groups (except physicians from which incomplete surveys were retained) were sampled randomly to construct groups of 46 each. The analytic sample was composed of 280 subjects. (Note that all analyses were repeated using the total sample and the results were similar in all substantial respects).

Results

The complete design was a 2 (value set, i.e., survey form; between subjects) \times 2 (facility size; between subjects) \times 6 (interest group; between subjects) \times 4 (program; within subjects) factorial. Because of an inadequate number of respondents in some cells of the full $2 \times 2 \times 6 \times 4$ design, we decided to follow a two-stage analysis strategy. The first stage tested for the presence of any instrument effects. The four different survey forms were constructed by combining the two different sets of 20 value statements (Value Sets A and B) with two sizes (small and large) of group homes. Because facility size was embedded in the description of programs, this first stage analysis was a 2 (value set; between subjects) \times 2 (size; between subjects) \times 4 (program; within subjects) factorial. For theoretical reasons, separate analyses were performed for attitudes and for behavioral intentions.

The facility size manipulation was not significant for either attitudes ($p > .05$; see Table 1) or behavioral intentions ($p > .05$). A significant effect for value set was found for attitudes ($p = .03$) but not for behavioral intentions ($p > .05$). Further analysis revealed that Value Set B elicited stronger expectancies (both more positively and negatively valenced) than did Value Set A; average value ratings were not significantly different between forms. Because neither facility size nor any two-way or three-way interactions involving facility size were significant, facility size was dropped from the second stage of analysis. Value set was retained in further analyses.

The second stage of analysis tested hypotheses generated by the receptivity model. A 2 (value set; between subjects) \times 4 (program; within subjects) \times 6 (interest group; between sub-

Table 1
Summary of Facility Size, Value Set, Program, and Interest Group Effects on Attitudes and Behavioral Intentions

Source	Attitudes				Behavioral intentions			
	<i>F</i>	<i>df</i>	<i>p</i>	η^2	<i>F</i>	<i>df</i>	<i>p</i>	η^2
Variables								
Facility size (S) ^a	0.0	1, 278	<i>ns</i>	—	0.0	1, 278	<i>ns</i>	—
Value set (V)	4.7	1, 278	.03	.03	0.4	1, 278	<i>ns</i>	—
Program (P)	49.4	3, 804	.01	.26	213.9	3, 804	.01	.62
Interest group (G)	2.7	5, 274	.02	.07	4.5	5, 274	.01	.07
Interactions								
S × V ^a	0.1	1, 276	<i>ns</i>	—	0.8	1, 276	<i>ns</i>	—
S × P ^a	0.7	3, 804	<i>ns</i>	—	1.5	3, 804	<i>ns</i>	—
V × P	0.5	3, 804	<i>ns</i>	—	1.1	3, 804	<i>ns</i>	—
V × G	1.7	5, 268	<i>ns</i>	—	.5	5, 268	<i>ns</i>	—
P × G	1.8	15, 804	.03	.05	1.3	15, 804	<i>ns</i>	—
S × V × P ^a	2.3	3, 828	.07	.01	.5	3, 828	<i>ns</i>	—
V × P × G	1.9	15, 804	<i>ns</i>	—	1.2	15, 804	<i>ns</i>	—
Covariates								
	<i>t</i>		<i>p</i>	η^2	<i>t</i>		<i>p</i>	η^2
Sex	0.89		<i>ns</i>	—	-0.03		<i>ns</i>	—
Age	2.66		.01	.04	1.32		<i>ns</i>	—
Education	0.53		<i>ns</i>	—	-0.39		<i>ns</i>	—
Socioeconomic status	0.12		<i>ns</i>	—	1.16		<i>ns</i>	—

^a Because facility size was included in only the Stage I analysis, all results which include facility size are from the 2 (size) × 2 (value set) × 4 (program) Stage I analysis. All other results are from the 2 (value set) × 4 (program) × 6 (interest group) Stage II analysis. Results from the Stage II analysis include the covariates.

jects) factorial analysis of variance (ANOVA) was performed to test whether attitudes and behavioral intentions varied by community interest group and by program. Sex, age, education, and SES were entered as covariates to investigate the predicted effects of interest groups distinct from the demographic characteristics of the members of these groups. SES was scaled using Hollingshead's (1957) two-factor index.

Attitudes

Value set ($p < .04$), interest group ($p < .03$), and program ($p < .001$) all had significant impacts upon attitudes. The only significant interaction was the interest group by program effect ($p < .03$). The *t* tests of the covariates indicated that only age ($p < .01$) was significant (see Table 1 for complete results).

Planned contrasts involving the following programs were conducted: (a) mental health group homes versus elderly group homes, $F(1, 279) = 18.75$, $p < .001$; (b) group homes versus outpatient care, $F(1, 279) = 0.05$, $p > .05$; and (c) deinstitutionalization programs (group home and outpatient care) versus state institutions, $F(1, 279) = 113.44$, $p < .001$. Respondents expressed significantly more favorable attitudes toward group homes for the elderly ($M = 33.64$) than similar group homes for mental health clients ($M = 26.46$). When both types of group homes and outpatient care were aggregated to compute a score for deinstitutionalization, it received a much more favorable attitude score ($M = 30.42$) than did state institutions ($M = 5.48$; see Table 2).

The interest group by program interaction affected only a few

comparisons. Regardless of program, vendors expressed the most favorable attitudes. Realtors indicated the least favorable attitudes toward all programs except state institutions. The remaining interest groups in descending order of overall attitude favorability were family members, state social service staff, government leaders, and physicians. The only exception to the above pattern of program results concerned the state social service staff. They indicated more favorable attitudes toward group homes for mental health clients than for elderly clients. They also favored outpatient care more than group home care (see Table 2).

Behavioral Intentions

Intentions of support were not significantly ($p > .05$) different across survey forms with different value sets. The type of program significantly affected intentions of support ($p < .001$). Significantly different intentions were expressed by the interest groups ($p < .001$). There were no significant interactions or covariates (see Table 1).

Again, planned contrasts were performed involving the following programs: (a) mental health group homes versus elderly group homes, $F(1, 279) = 330.9$, $p < .001$; (b) group home care versus outpatient care, $F(1, 279) = 93.06$, $p < .001$; and (c) deinstitutionalization programs versus state institutions, $F(1, 279) = 246.11$, $p < .001$. Respondents reported more intentions of support for elderly group homes ($M = 1.58$) than for group homes for mental health clients ($M = 0.10$). They were more supportive of outpatient care ($M = 1.42$) than of group home

Table 2
Mean Attitude Scores for Programs and Interest Groups

Interest group	Programs					Overall interest-group mean
	Retarded group homes ^a	Mental health group homes	Elderly group homes	Outpatient care	State institution	
Government leaders (1)	38.58	29.00	41.20	33.83	12.22	24.97
Home vendors (2)	36.37	45.59	51.50	50.02	10.83	38.86
Families of clients (3)	47.56	27.56	41.41	28.35 ^b	10.70	31.11
Realtors (4)	10.83	11.02 ^b	26.87 ^b	13.58 ^c	0.89	12.64
Physicians (5)	14.86	17.47 ^b	27.91 ^b	26.66 ^b	-4.11	16.56
State social service staff (6)	25.63	31.84	30.35 ^b	34.74	10.41	26.59
Overall program mean	29.39	27.69	36.90	31.38	7.28	26.53

^a Means for mental retardation group homes are based on data from facilities described as large in size and are included only for illustrative purposes. This program was not included in the reported analyses. ^b Significant differences among groups = 2. ^c Significant differences among groups = 1, 2, 6.

residential care ($M = 0.84$). Finally, respondents intended to oppose state institutions ($M = -0.43$) but intended to support deinstitutionalization programs ($M = 1.04$).

All interest groups indicated intentions to oppose state institutions. Group home care for mental health clients was the least supported of the deinstitutionalization programs and was opposed by realtors and physicians. All groups except vendors and families of clients expressed their highest level of support for group home care for the elderly. Vendors and families of clients, however, indicated stronger intentions of support for outpatient care. Vendors expressed significantly more support in general than realtors, physicians, and families of clients (see Table 3 for complete results).

Finally, the correlations between attitudes toward each program and corresponding behavioral intention ratings were significant ($p < .001$): for group homes serving (a) the mentally retarded, $r = .37$, (b) mental health clients, $r = .39$, and (c) the

elderly, $r = .34$; for outpatient programs, $r = .40$; for state institutions, $r = .37$; and for all deinstitutionalization programs combined as a policy, $r = .45$.

Conclusions

The results strongly supported our hypotheses. Only facility size failed to have the predicted effect.

All community groups strongly preferred deinstitutionalization programs over institutionalization. Mean receptivity (i.e., attitudes) ratings were as much as four times greater, and mean intention scores as much as six times more supportive for deinstitutionalization programs than for institutionalization.

Program type and interest group influenced both attitudes and intentions. Group homes for the elderly received more positive attitudes and more support than group homes for the mentally ill. Attitudes for outpatient programs were not different

Table 3
Mean Behavioral Intention Scores for Programs and Interest Groups

Interest group	Programs					Overall interest-group mean
	Retarded group homes ^a	Mental health group homes	Elderly group homes	Outpatient care	State institution	
Government leaders (1)	0.85	0.00 ^b	1.85	1.39 ^b	-0.30	0.76
Home vendors (2)	1.79	0.85	2.11	2.15	-0.33	1.31
Families of clients (3)	0.56	0.02	1.35 ^c	1.57	-0.02	0.70
Realtors (4)	0.25	-0.37 ^b	1.15 ^d	1.00 ^c	-0.50	0.31
Physicians (5)	0.07	-0.54 ^b	1.14 ^d	0.69 ^e	-0.66	0.15
State social service staff (6)	1.13	0.20	2.04	1.91	-0.50	0.96
Overall program mean	0.80	0.05	1.63	1.48	-0.37	0.72

Note. Behavioral intention score range = -3 to 3.

^a Means for mental retardation group homes are based on data from facilities described as large in size and are included only for illustrative purposes. This program was not included in the reported analyses. ^b Significant differences among groups = 2. ^c Significant differences among groups = 2, 6. ^d Significant differences among groups = 1, 2, 6. ^e Significant differences among groups = 1, 2, 3, 6.

from those for group homes but intentions of support were greater for outpatient programs than for group homes. Realtors had the least favorable attitudes and least supportive intentions of all interest groups. Group home operators (i.e., vendors) had the most favorable. Only the state social service staff expressed a preference pattern different from the other groups; they preferred outpatient programs over group homes and preferred group homes for the mentally ill over those for the elderly.

Of the demographic variables, only age was significant and only for attitudes. Interest group membership accounted for twice as much variation in attitudes as age (see Table 1). Interest group, but not demographic variables, influenced intentions of support. However, the results of this study should not be considered representative of the interest groups sampled because of the low response rates. Without knowing such demographic information as mean age or sex proportions for each interest group population, it was not possible to test for response bias. We assumed similar self-selection forces operated in all groups. The differences between groups were relevant to testing our model given this assumption.

The interest group variable accounted for approximately the same amount of variation in both attitudes and intentions. The type of program, however, explained twice as much variation in intentions as in attitudes and was from three to ten times more influential than the interest group variable (Table 1). Additional research is needed to determine why the type of program affects intentions more than attitudes.

We did not investigate the relation between program distance from the respondent's home and attitudes and intentions. Although we expected a relation between these variables, we assumed that distance did not influence the relation between the variables tested in our model.

The results substantially supported our model of community receptivity based on individual attitudes and interest group membership. Attitudes, however, only moderately correlated with intentions. One post-hoc explanation is that attitudes are only one component of intentions. We did not assess the costs and benefits of carrying out an intention. Such costs might include unwanted personal publicity and loss of time from more pleasant activities. Our present attitude model of receptivity did not include this additional motivational component of intentions. We propose that Lawler's (1973) expectancy theory of motivation be incorporated with our expectancy-value model of receptivity. Additional variables that might intervene between attitudes and intentions need to be further explored.

Our evidence supports an interest group approach over simply using demographic variables for explaining receptivity. Although interest group membership accounted for a small amount of variation, it has heuristic value for planning attitude change efforts because it permits more focused targeting than using demographic variables. Used judiciously, information campaigns could potentially both (a) fill knowledge gaps (lack of awareness) and (b) correct false beliefs (expectancies) about a program (Pratt & Kethley, 1980). Misperceptions (factually incorrect expectancies) accounted for about one-half of the reasons involved in attitudes unrelated to deinstitutionalization (Severy, 1984). Clearly, inadequacies exist in some deinstitutionalization programs (Lamb, 1979) and some residents' fears,

such as the dangerousness of clients (Steadman, 1981), may be realistic.

Because we did not actually compare receptivity attitudes with postsurvey actions of support or opposition, we can only cite similar applications in other contexts to support the predictive validity of our approach. Scheirer (1981) found that how much an organization's staff liked (valued) aspects of a program and how much they believed (expectancy) in the efficacy of the new program were both independently related to the actual amount of staff implementation. The multiplicative product measure of attitudes, rather than values and expectancies separately, better predicted family planning behavior over a 9-month period (Severy, 1982).

Other assessment procedures do not provide the detailed information about the values and beliefs for support or opposition that our procedure does (Hagedorn, Beck, Neubert, & Werlin, 1976; Johnson & Beditz, 1981; Miller, 1981; Shadish et al., 1982). Only our procedure used the more predictive multiplicative scoring (for examples of nonmultiplicative scoring see Scheirer, 1981; Sundstrom et al., 1977).

Sigelman (1976) has argued that assessing receptivity for group homes was unnecessary because (a) only a minority of residents were aware of a mental health facility in their neighborhood (Heinemann, Perlmutter, & Yudin, 1974; Morrison & Libow, 1977) and (b) in most instances, residents who opposed these facilities changed their attitudes after the facility's operation began (Baker & Seltzer, 1977; Kennedy, Felner, Blank, & Eisenstat, 1982; Smith, 1981). However, in locations such as Florida where group homes must get zoning approval prior to construction or operation, receptivity before implementation is crucial. Public notices required for rezoning permits can alert and mobilize a community.

Local government leaders as a group were supportive of group homes and outpatient care. In practice, however, attention must be paid to specific leaders in particular communities rather than to leaders as a group. Leaders are more likely to act if local citizens (as members of organized or unorganized interest groups) publicly express their support. With the exception of realtors, widespread but latent support was found for deinstitutionalization programs. Passive receptivity (i.e., absence of protest) is often inadequate for successful implementation and thus latent support must be organized into public action. Actual community or organizational decisions about planned change require intergroup models of influence, power and decision making (O'Brien, 1975). Although we have presented our model as a general model of receptivity toward change, it obviously does not address numerous variables related to change found by previous research (Rogers & Shoemaker, 1971).

We have investigated only receptivity toward change prior to the decision to implement at the local level. Implementation success, however, depends on a series of actions at successive points in time. Each action may be dependent on different sets of actors, values, and/or expectancies (Pelz, 1981a, 1981b). Although our procedure has been applied in other settings, at different points in the implementation process (Severy, Houlden, & Wilmoth, 1979; Severy, Houlden, Wilmoth, & Silver, 1982; Severy & Whitaker, 1982), future research is needed to

further explore the relation between receptivity and implementation cycle. Our approach, consistent with the social systems approach to change (Shadish, 1984), may also be applicable to organizational change as well as to community change.

References

- Anderson, S. B., & Bell, S. (1978). *The profession and practice of program evaluation*. San Francisco: Jossey-Bass.
- Baker, B., & Seltzer, M. (1977). *As close as possible: Community residences for retarded adults*. Boston: Little-Brown.
- Baron, R., & Piasecki, J. (1981). The community versus community care. In R. Budson (Ed.), *New directions for mental health services: Issues in community residential care* (pp. 63-76). San Francisco: Jossey-Bass.
- Braun, P., Kochansky, G., Shapiro, R., Greenberg, S., Gudeman, J., Johnson, S., & Shore, M. (1981). Overview: Deinstitutionalization of psychiatric patients, a critical review of outcome studies. *American Journal of Psychiatry*, 138, 736-749.
- Brockman, J., D'Arcy, C., & Edmonds, L. (1979). Facts or artifacts: Changing public attitudes toward the mentally ill. *Social Science and Medicine*, 13, 673-682.
- Calder, B. J. (1977). Focus groups and the nature of qualitative marketing research. *Journal of Marketing Research*, 14, 353-364.
- CBS (1980). *60 Minutes: Not in my neighborhood*. [Television]. New York.
- Dean, M. (1977). Impact of mental health facilities on property values. *Community Mental Health Journal*, 13, 150-157.
- Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior*. Reading, MA: Addison-Wesley.
- Fracchia, J., Canale, D., Cambria, E., Ruest, E., & Sheppard, C. (1976). Public views of ex-mental patients: A note on perceived dangerousness and unpredictability. *Psychological Reports*, 38, 495-498.
- Hagedorn, H. J., Beck, K. J., Neubert, S. F., & Werlin, S. H. (1976). *A working manual of simple program evaluation techniques for community mental health centers*. Rockville, MD: National Institute of Mental Health.
- Heinemann, S., Perlmutter, F., & Yudin, L. (1974). The community mental health center and community awareness. *Community Mental Health Journal*, 10, 221-227.
- Hollingshead, A. (1957). *Two-factor index of social position*. New Haven, CT: Yale University Press.
- Johnson, P., & Beditz, J. (1981). Community support systems: Scaling community acceptance. *Community Mental Health Journal*, 17, 153-160.
- Kennedy, M., Felner, R., Blank, M., & Eisenstat, R. (1982, March). *Neighborhood attitudes toward community-based, residential care: Changes over time*. Paper presented at the meeting of the Southeastern Psychological Association, New Orleans, LA.
- Kiesler, C. (1982). Mental hospitals and alternative care: Noninstitutionalization as potential public policy for mental patients. *American Psychologist*, 37, 349-360.
- Lamb, H. R. (1979). The new asylums in the community. *Archives of General Psychiatry*, 36, 129-138.
- Lawler, E. (1973). *Motivation in work organizations*. Monterey, CA: Brooks/Cole.
- Locker, D., Rao, B., & Weddell, J. (1979). The community reaction to a hostel for the mentally handicapped. *Social Science and Medicine*, 13, 817-821.
- Maypole, D. (1981). Fears about the development of a group home. *Administration in Mental Health*, 9, 67-75.
- Mazmanian, D., & Sabatier, P. (1980). The role of attitudes and perceptions in policy evaluation by attentive elites: The California Coastal Commissions. In H. Ingram & D. Mann (Eds.), *Why policies succeed or fail* (pp. 107-133). Beverly Hills, CA: Sage.
- Miller, F. (1981). Mental health center versus community perceptions of mental health services. *Journal of Community Psychology*, 9, 204-209.
- Morrison, J., & Libow, J. (1977). The effect of newspaper publicity on a mental health center's community visibility. *Community Mental Health Journal*, 13, 58-62.
- Morrison, J. K., Smith, J. E., Fentiman, J. R., Madrazo-Peterson, R., and Boyagian, P. S. (1979). Attitudes of community gatekeepers and psychiatric social workers toward mental illness. *Journal of Community Psychology*, 7, 147-150.
- Murphy, J., & Datel, W. (1976). A cost-benefit analysis of community versus institutional living. *Hospital and Community Psychiatry*, 27, 165-170.
- O'Brien, D. (1975). *Neighborhood organization and interest group processes*. Princeton, NJ: Princeton University Press.
- Pelz, D. (1981a, August). *Information use in urban innovations: Actors and channels*. Paper presented at the American Psychological Association, Los Angeles, CA.
- Pelz, D. (1981b, October). "Staging" effects in adoption of urban innovations. Paper presented at the meeting of the Evaluation Research Society, Austin, TX.
- Pratt, C., & Kethley, A. (1980). Anticipated and actual barriers to developing community mental health programs for the elderly. *Community Mental Health Journal*, 16, 205-216.
- Rabkin, J. (1972). Opinions about mental illness: A review of the literature. *Psychological Bulletin*, 77, 153-171.
- Rabkin, J. (1975). The role of attitudes toward mental illness in evaluation of mental health programs. In M. Guttentag & E. Struening (Eds.), *Handbook of evaluation research* (pp. 431-482). Beverly Hills, CA: Sage.
- Rogers, E., & Shoemaker, F. (1971). *Communication of innovations: A cross cultural approach*. New York: Free Press.
- Rosenberg, M. (1956). Cognitive structure and attitudinal affect. *Journal of Abnormal and Social Psychology*, 53, 367-373.
- Rothbart, M. (1973). Perceiving social injustice: Observations on the relationship between liberal attitudes and proximity to social problems. *Journal of Applied Social Psychology*, 3, 291-302.
- Scheirer, M. (1981). *Program implementation: The organizational context*. Beverly Hills, CA: Sage.
- Segal, S., Baumohl, J., & Moyses, E. (1980). Neighborhood types and community reaction to the mentally ill: A paradox of intensity. *Journal of Health and Social Behavior*, 21, 345-359.
- Severy, L. (1982). *Contraceptive decisions: Spousal relationships, method commitment and expectancy-values*. (Report No. NO1-HD-02802). Washington, DC: Center for Population Research, National Institute of Child Health and Human Development.
- Severy, L. (1984). Research on reproductive behavior in developed and developing countries. *Proceedings of the 23rd International Congress of Psychology*. Amsterdam: North Holland Publishing.
- Severy, L., Houlden, P., & Wilmoth, G. (1979). *Community attitudes toward community-based treatment and restitution programs* (Tech. Rep. STAR-78-1141). Gainesville: University of Florida, Department of Psychology.
- Severy, L., Houlden, P., & Wilmoth, G. (1981). Community acceptance of innovative programs. In L. Bickman (Ed.) *Applied Social Psychology Annual* (Vol. 2, pp. 71-95). Beverly Hills, CA: Sage.
- Severy, L., Houlden, P., Wilmoth, G., & Silver, S. (1982). Community receptivity to juvenile justice program planning. *Evaluation Review*, 6, 25-46.
- Severy, L., & Whitaker, J. (1982). Assessment of interagency relation-

- ships: A quantitative process analysis and expectancy-value approach. *Evaluation Review*, 6, 267-277.
- Shadish, W. (1984). Policy research: Lessons from the implementation of deinstitutionalization. *American Psychologist*, 39, 725-738.
- Shadish, W., Thomas, S., & Bootzin, R. (1982). Criteria for success in deinstitutionalization: Perceptions of nursing homes by different interest groups. *American Journal of Community Psychology*, 10, 553-566.
- Sigelman, C. (1976). A Machiavelli for planners: Community attitudes and selection of a group home site. *Mental Retardation*, 34, 26-29.
- Smith, C. (1981). Hospital proximity and public acceptance of the mentally ill. *Hospital and Community Psychiatry*, 32, 178-180.
- Steadman, H. (1981). Critically reassessing the accuracy of public perceptions of the dangerousness of the mentally ill. *Journal of Health and Social Behavior*, 22, 310-316.
- Stone, C. (1980). The implementation of social programs: Two perspectives. *Journal of Social Issues*, 36, 13-34.
- Sundstrom, E., Lounsbury, J., Schuller, C. R., Fowler, J., & Mattingly, T. (1977). Community attitudes toward a proposed nuclear power generating facility as a function of expected outcomes. *Journal of Community Psychology*, 5, 199-208.
- Taylor, S., Dear, M., & Hall, G. (1979). Attitudes toward the mentally ill and reactions to mental health facilities. *Social Science and Medicine*, 13D, 281-290.
- Test, M. A. (1981). Effective community treatment of the chronically mentally ill: What is necessary. *Journal of Social Issues*, 37, 71-86.
- Tringo, J. L. (1970). The hierarchy of preference toward disability groups. *Journal of Special Education*, 4, 295-306.
- Walton, R. E. (1975). The diffusion of new work structures: Explaining why success didn't take. *Organizational Dynamics*, 3(3), 3-22.
- Weisbrod, B., Test, M., & Stein, L. (1980). Alternatives to mental hospital treatment: Economic benefit-cost analysis. *Archives of General Psychiatry*, 37, 400-405.
- Wilgus, A., & Epstein, I. (1978). Group homes for adolescents: A comparative case study. *Social Work*, 23, 486-491.

Received February 12, 1986 ■

SHORT NOTES

Averaging Correlation Coefficients: Should Fisher's z Transformation Be Used?

N. Clayton Silver and William P. Dunlap
Tulane University

Averaging correlations leads to underestimation because the sampling distribution of the correlation coefficient is skewed. It is also known that if correlations are transformed by Fisher's z prior to averaging, the resulting average overestimates the population value of z . The behavior of these procedures for averaging correlations was investigated via Monte Carlo simulation, both in terms of bias (under- and overestimation) and precision (standard errors). It was found that average z backtransformed to r is less biased positively than average r is biased negatively. The standard error of average r was smaller than that of average z when the population correlation was small; however, the reverse was true when the population correlation exceeded .5. Regardless of sample size, backtransformed average z was always less biased; therefore, the use of the z transformation is recommended when averaging correlation coefficients, particularly when sample size is small.

The correlation coefficient is such a notoriously variable statistic that large sample sizes are required to get stable estimates. There are instances, however, in which small sample sizes are used and the individuals are tested repeatedly (Carter, Kennedy, & Bittner, 1981; Mackaman, Bittner, Harbeson, Kennedy, & Stone, 1982). It has been shown by Dunlap, Jones, and Bittner (1983) and Dunlap, Silver, Hunter, and Bittner (1985) that estimates of correlations with small samples are improved dramatically by averaging correlations over repeated measures of the variables involved.

If one is to average correlations in an attempt to get more stable estimates, the question remains of how best to average correlations. First, one could convert r to Fisher's z transformation, average the z s, and backtransform to r (e.g., Rambo, Chomiak, & Price, 1983), or one could use a simple arithmetic average, in which case the correlations would be summed and divided by the number of coefficients (e.g., Turnage & Muchinsky, 1984). There is, however, a problem of bias with the latter method. Because the distribution of r becomes negatively skewed as the correlation is larger than zero, the average r tends to underestimate the population correlation. An expression for this bias in average r (adapted from Kendall & Stuart, 1979) is

$$E(r - \rho) = -\rho(1 - \rho^2)/2N + O(1/N^2), \quad (1)$$

where $O(1/N^2)$ stands for other terms not larger than $1/N^2$. Yet, Schmidt, Gast-Rosenberg, and Hunter (1980) have indicated that it is best to use average r , because they felt that bias in aver-

aging r is negligible except for very large values of r , and to the extent bias exists, their findings would be conservative.

Fisher's (1921) z transformation almost entirely corrects the skew in the distribution of r , where

$$z = 0.5 \log_e[(1 + r)/(1 - r)]. \quad (2)$$

Because the distribution of z is approximately normal, one might expect averaged z s to be less biased than averaged r s. It turns out, however, that whereas averaged r s are negatively biased, averaged z s are positively biased. This expression (again adapted from Kendall & Stuart, 1979) is

$$E(z - \zeta) = \rho/[2(N - 1)] + O(1/N^2). \quad (3)$$

What is not known is the statistical behavior of an averaged z , backtransformed to the form of r . The equation for the backtransformation is

$$r = (e^{2z} - 1)/(e^{2z} + 1). \quad (4)$$

The present article, therefore, used a Monte Carlo simulation to compare two procedures for averaging correlations. The first was to compute the simple arithmetic average of correlation coefficients; the second was to transform r to z prior to averaging, average the z s, then convert the average z back to the form of r .

Method

The first step in the Monte Carlo simulation was to generate pseudo-random normal deviates via Box and Muller's (1958) procedure using random fractions generated by the Dec-20 internal random number function, RAN (see Edgell, 1979, for the characteristics of this function). A matrix of weightings was derived that when applied to the ran-

Correspondence concerning this article should be addressed to William P. Dunlap, Department of Psychology, Tulane University, 2007 Percival Stern Hall, New Orleans, Louisiana 70118.

Table 1
Means and Standard Errors of Averaged *r* and Backtransformed *z* With Sample Sizes of 10, 20, and 30

<i>n</i> and statistic		Population correlation									
		.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
10											
<i>r</i>	<i>M</i>	−.004	.093	.193	.283	.382	.475	.576	.681	.784	.887
	<i>SE</i>	.093	.119	.140	.154	.160	.163	.151	.132	.103	.065
<i>z</i>	<i>M</i>	−.005	.103	.210	.305	.407	.501	.602	.704	.802	.897
	<i>SE</i>	.103	.130	.150	.161	.165	.165	.149	.128	.098	.061
20											
<i>r</i>	<i>M</i>	.001	.097	.194	.292	.390	.493	.591	.694	.793	.894
	<i>SE</i>	.063	.085	.100	.106	.112	.109	.103	.084	.066	.037
<i>z</i>	<i>M</i>	.001	.102	.202	.303	.402	.505	.603	.704	.800	.899
	<i>SE</i>	.069	.089	.103	.108	.113	.109	.102	.082	.065	.036
30											
<i>r</i>	<i>M</i>	.000	.098	.197	.298	.395	.495	.589	.694	.795	.897
	<i>SE</i>	.052	.068	.080	.086	.088	.085	.082	.069	.052	.029
<i>z</i>	<i>M</i>	.000	.101	.202	.304	.403	.504	.597	.701	.800	.900
	<i>SE</i>	.054	.070	.082	.088	.089	.085	.082	.069	.051	.028

dom independent normal deviates, produced new variables as linear composites of the independent random normal variables that had the appropriate population correlation matrix (Dunlap et al., 1985).

In this case, the population intercorrelation matrix had all correlations equal to a specified value. Intercorrelations between all variables were computed and averaged as they were, or first converted by Fisher's *z*, averaged, then backtransformed via Equation 4.

Means and standard errors for both average *r* and backtransformed average *z* were calculated and accumulated for each data generation. The cycle was repeated 10,000 times for each data point.

Results and Discussion

The empirical estimates of the two methods of averaging can be seen in Table 1 for average *r* and backtransformed average *z*, respectively, and the corresponding standard errors.

Preliminary inspection of the data showed that the number of replications, which determined the number of correlations averaged each time, had only minor effects on the extent of bias seen for the two correlation averaging procedures. For this reason, the findings were averaged across the replications variable and are presented in Table 1 as functions of the sample size, *n*, and the population correlation.

Bias

As can be seen in Table 1, the backtransformed average *z* values, in most cases, fell slightly above the correct values, whereas average *r* tended to fall below, and in the case of intermediate correlations, substantially below the true population correlation. The bias, or deviation of the estimated value from the population value is depicted in Figure 1. Here it can be seen that the negative bias in average *r* is substantially greater than the positive bias of backtransformed average *z*, particularly at

the smaller sample sizes. With a sample size of 30, although there is still greater bias in average *r* than in average *z*, the absolute amount of bias in both estimators is fairly small. Because bias or accuracy of an estimation procedure is of first order importance, these findings clearly support the use of backtransformed average *z* as opposed to average *r* when sample sizes are small. When sample sizes are much above 30, because the bias in both procedures is small, it is less clear whether the extra effort required to convert to Fisher's *z* prior to averaging is worth the trouble.

Variability

The standard errors of the two types of estimates are shown in Table 1. As can be seen, the standard errors cross at a popula-

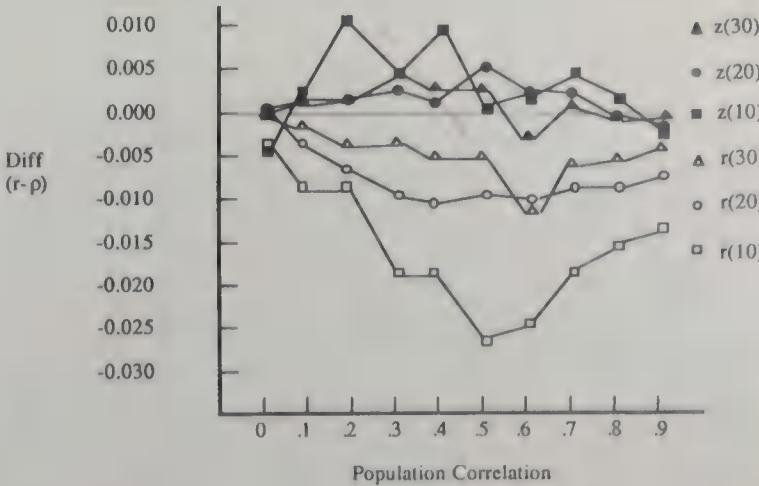


Figure 1. Bias (difference between estimated and actual population correlation) as functions of the population correlation and sample size for average *r* and backtransformed average *z*.

tion correlation of about .5, such that average r shows less variability for the lower correlations, whereas average z is more precise for larger correlations. Again, it is clear from Table 1 that the differences in standard errors are negligible with a sample size of 30, so that this secondary factor is of importance only for a small sample.

Therefore, we concluded that there is substantial benefit to be gained by transforming correlation coefficients to Fisher's z prior to averaging, then backtransforming the average, especially if sample sizes are small. Although it appears that this procedure will always be less biased than simply averaging correlations, from a practical standpoint the amount of bias when the sample size is large is so small that we question the necessity of the extra effort involved. It was somewhat surprising to find a crossover in the standard error functions of the two averaging procedures; however, compared to the primary importance of bias, these differences in precision are not great and are certainly of secondary importance.

These findings lent empirical support to the procedure of Dunlap et al. (1983, 1985), who used backtransformed average z rather than average r when performing Monte Carlo simulations for studying gains in precision that could be accomplished by averaging correlations. We hope that these results will serve as a guide to future researchers who wish to average correlation coefficients and who recognize that bias may be a problem, particularly with smaller sample sizes.

References

- Box, G. E. P., & Muller, N. E. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, 29, 610-611.
- Carter, R. C., Kennedy, R. S., & Bittner, A. C., (1981). Grammatical reasoning: A stable performance yardstick. *Human Factors*, 23, 587-591.
- Dunlap, W. P., Jones, M. B., & Bittner, A. C. (1983). Average correlations vs. correlated averages. *Bulletin of the Psychonomic Society*, 21, 213-216.
- Dunlap, W. P., Silver, N. C., Hunter, R. E., & Bittner, A. C. (1985). Averaged cross-correlations: A methodology for validity assessment in small samples. In R. Eberts, & C. G. Eberts (Eds.), *Trends in ergonomics/human factors* (Vol. 2, pp. 13-21). Amsterdam: Elsevier.
- Edgell, S. E. (1979). A statistical check of the DECsystem-10 FORTRAN pseudorandom number generator. *Behavior Research Methods & Instrumentation*, 11, 529-530.
- Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1, 1-32.
- Kendall, M., & Stuart, A. (1979). *The advanced theory of statistics* (Vol. 1, 4th ed.). New York: Macmillan.
- Mackaman, S. L., Bittner, A. C., Harbeson, M. M., Kennedy, R. S., & Stone, D. A. (1982). Performance evaluation tests for environmental research (PETER): Wonderlic Personnel Test. *Psychological Reports*, 51, 635-644.
- Rambo, W. W., Chomiak, A. M., & Price, J. M. (1983). Consistency of performance under stable conditions of work. *Journal of Applied Psychology*, 68, 78-87.
- Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J. E. (1980). Validity generalization results for computer programmers. *Journal of Applied Psychology*, 65, 643-661.
- Turnage, J. J. & Muchinsky, P. M. (1984). A comparison of the predictive validity of assessment center evaluations versus traditional measures in forecasting supervisory job performance: Interpretive implications of criterion distortion for the assessment paradigm. *Journal of Applied Psychology*, 69, 595-602.

Received December 9, 1985

Revision received June 26, 1986 ■

Reexamining the Component Stability of Owens's Biographical Questionnaire

Gary J. Lautenschlager and Garnett Stokes Shaffer
University of Georgia

The component stability of Owens's Biographical Questionnaire (BQ) is reexamined in light of a reanalysis of the original data used to develop the 118-item BQ. Results indicate that components for men are stable over time and geographical location. In addition, the stability of a set of components for women is now demonstrated. The present results suggest that the differences found by Eberhardt and Muchinsky (1982) in the components obtained from women are due to methodological differences between studies and not to changing life experiences of women over the time period between the studies. The components for women that were obtained from a reanalysis of the original data used by Owens and Schoenfeldt (1979) do replicate.

The component stability of Owens's Biographical Questionnaire (BQ) was recently investigated by Eberhardt and Muchinsky (1982). They reported finding essentially the same set of components for men as had been obtained by Owens and Schoenfeldt (1979), but there were marked differences in comparing components for women. Although several plausible explanations for the discrepancies in the components obtained for women were mentioned, Eberhardt and Muchinsky found the most compelling explanation to be that of the changing life experiences of women over time.

The present study was undertaken to determine if the lack of similarity for the women's components obtained by Eberhardt and Muchinsky (1982) might be due to methodological differences in the ways that the components were obtained. Owens and Schoenfeldt (1979) described the intricate and rather elaborate process of moving from a pool of 2,000 items to the final 118-item BQ form. It is important to note that the components retained and interpreted in that study and the associated item loadings were those based on a pool of 275 items. Eberhardt and Muchinsky had only the 118-item BQ and derived components from these items. The difference in the number of items used in the component analyses (i.e., 118 vs. 275) was mentioned by Eberhardt and Muchinsky, but they did not believe that this was a serious limitation. In addition, they did not have the complete item-loading information available to explore this possibility in detail.

Method

Procedure

The components that were interpreted and used for final item selection in the original development of the 118-item Owens's BQ were actu-

ally drawn from principal components analyses, done separately for each sex, that each involved a total of 275 items (Owens & Schoenfeldt, 1979). In the present study, only the data for the 118 items that appeared in the final form of the BQ were used in each principal components analysis of the original data (Owens & Schoenfeldt, 1979). Following the procedure used by Eberhardt and Muchinsky (1982), separate principal component analyses were performed for men and for women. A total of 13 components were extracted for the men and 15 components for the women, using the SAS factor procedure (Statistical Analysis System; SAS Institute, 1982). The obtained components were then rotated using the varimax criterion (Kaiser, 1958).

Measures of Component Similarity

Three indices of component similarity were used to compare the Eberhardt and Muchinsky (1982) components with the components obtained by reanalyzing the original data. The first two indices used were the percentage overlap measure and the correlation of between-study component loadings (Eberhardt & Muchinsky, 1982, p. 139). These were the measures used in the Eberhardt and Muchinsky study. In addition, the coefficient of factor congruence (Burt, 1948; Tucker, 1951; Wrigley & Neuhaus, 1955) was also used. This measure is similar to the correlation coefficient, but does not obscure mean differences in loadings.

Results

Comparisons were made between the components obtained in the present reanalysis of Owens's data with components obtained by Eberhardt and Muchinsky (1982) and also with the components reported in Owens and Schoenfeldt (1979). Tables 1 and 2 present the resulting components derived from the reanalysis of Owens's data for men and women, respectively. In these tables component labels are used to indicate the predominant item content for the components obtained in the reanalysis of Owens's data. The item content of the components closely corresponds to that reported in earlier investigations, and so is not repeated here (cf. Eberhardt & Muchinsky, 1982; Owens & Schoenfeldt, 1979; Schoenfeldt, 1974). The proportion of total variance accounted for by the 13 components for men in the present study was 39.7%, compared to 45.6% for the compo-

Portions of this article were presented at the annual meeting of the American Psychological Association, Washington, DC, August 1986.

The authors wish to thank Bruce J. Eberhardt for providing component loadings, and Barry Blakley for his assistance with data entry.

Correspondence concerning this article should be addressed to Gary J. Lautenschlager, Department of Psychology, University of Georgia, Athens, Georgia 30602.

Table 1
Comparison of Component Overlap Across Studies for Men

Component ^a	Component comparison				
	With EM		With OS		EM & OS % overlap
	Component no.	% overlap	Component no.	% overlap	
1. Academic achievement	2	88	2	97	85
2. Athletic interest	1	91	4	95	95 ^b
3. Warmth of parental relationship	9	96	1	92	96 ^b
4. Socioeconomic status	4	100	7	95	95
5. Social introversion	5	76	3	96	73 ^b
6. Aggressiveness/independence	11	80	6	87	84 ^b
7. Intellectualism	10	82	5	95	86
8. Scientific interest	8	83	10	78	88 ^b
9. Parental control vs. freedom	3	96	8	96	100
10. Positive academic attitude	13	80	11	94	84
11. Social desirability	7	82	9	76	95
12. Religious activity	6	89	12	91	80
13. Sibling friction	12	100	13	100	100

Note. EM = Eberhardt & Muchinsky (1982); OS = Owens & Schoenfeldt (1979).
^a Obtained from reanalysis of original Owens's data.
^b These overlap values do not agree with those presented by Eberhardt & Muchinsky (1982). Details describing the differences are available upon request from the first author.

nents from Eberhardt and Muchinsky. The 15 components for women accounted for 46.4% of the total variance in the present study, compared to 49.6% for the components from Eberhardt and Muchinsky. Because the method of principal components analysis was used, it is most appropriate to discuss each component's contribution to total variance. Common variance, as discussed by Eberhardt and Muchinsky (1982), has more significance in the context of a common factor analysis model.

Tables 1 and 2 also present the percentage overlap figures for the most similar components across studies. In Table 1 it is clear that for men the components exhibit considerable overlap across analyses, and thus were unaffected by use of a smaller item pool (118 vs. 275 items) in the reanalysis. All 13 components for men overlap considerably across all three studies, with values ranging from 73% to 100% overlap. However, the results in Table 2 indicate that the use of the

Table 2
Comparison of Component Overlap Across Studies for Women

Component ^a	Component comparison				
	With EM		With OS		EM & OS % overlap
	Component no.	% overlap	Component no.	% overlap	
1. Academic achievement	2	88	3	92	96
2. Athletic participation	5	100	7	94	94 ^b
3. Social leadership	1	79	2	57	61
4. Parental control	3	95	5	100	95 ^b
5. Socioeconomic status	4	100	4	95	95
6. Warmth of paternal relationship	10	89	1	70	70
7. Independence/dominance	13	76		No comparable component	
8. Scholastic and cultural interest	11	72		No comparable component	
9. Scientific interest	12	91	8	78	78 ^b
10. Positive academic attitude	14	89	14	88	88 ^b
11. Religious activity	6	100		No comparable component	
12. Sibling friction	9	75		No comparable component	
13. Warmth of maternal relationship	8	67	1	67	56 ^b
14. Negative social adjustment	7	72		No comparable component	
15. Approachability	—			No comparable component	—

Note. EM = Eberhardt & Muchinsky (1982); OS = Owens & Schoenfeldt (1979).
^a Obtained from reanalysis of original Owens's data.
^b These overlap values do not agree with those presented by Eberhardt & Muchinsky (1982). Details describing the differences are available upon request from the first author.

Table 3
Summary of Maximum Between-Study Component Similarity Measures for Men

Component		Component similarity measure	
Reanalysis of Owens' data	Eberhardt & Muchinsky (1982)	Congruence	Correlation
1. Academic achievement (5.75; 14)	2. (7.04; 18)	.77	.75
2. Athletic interest (5.30; 11)	1. (5.63; 11) ^a	.93	.92
3. Warmth of parental relationship (4.21; 11)	9. (4.71; 12)	.93	.92
4. Socioeconomic status (4.07; 11)	4. (4.10; 11)	.93	.93
5. Social introversion (4.03; 12)	5. (4.73; 17)	.82	.83
6. Aggressiveness/independence (3.92; 12)	11. (3.00; 8)	.73	.67
7. Intellectualism (3.80; 11)	10. (3.16; 11)	.67	.64
8. Scientific interest (3.79; 11)	8. (3.94; 13)	.68	.65
9. Parental control vs. freedom (3.76; 12)	3. (4.54; 11)	.89	.89
10. Positive academic attitude (2.93; 9)	13. (3.40; 11)	.81	.78
11. Social desirability (2.93; 11)	7. (3.56; 11)	.75	.73
12. Religious activity (2.65; 5)	6. (3.27; 4)	.87	.87
13. Sibling friction (2.59; 5)	12. (2.77; 5)	.85	.84

Note. All correlations are significant ($p < .05$). The numbers in parentheses next to each component represent the eigenvalue and the number of items that load on each component, respectively.

^a An examination of the loadings obtained from Eberhardt showed 11 items with loadings greater than .30 in absolute value on this component, rather than the 10 cited in Eberhardt and Muchinsky (1982).

smaller 118-item pool in the reanalysis produced a set of components for women that show the greatest overlap with the components obtained by Eberhardt and Muchinsky (1982). Whereas Eberhardt and Muchinsky reported that only 9 of their components for women had significant overlap (defined as 60% or more overlap) with the Owens and Schoenfeldt (1979) female components, 14 out of the 15 components for women obtained in the reanalysis did overlap with the Eberhardt and Muchinsky components. The overlap values for these 14 components range from 67% to 100% overlap. Note that essentially the same components that were not replicated in the Eberhardt

and Muchinsky study also failed to occur in the reanalysis of the original data.

Further comparisons between components focuses mainly on the components obtained in the reanalysis of the original data with those obtained by Eberhardt and Muchinsky (1982). Table 3 presents the maximum between-study correlations and congruence values for the components obtained for men. In addition, information is also presented regarding the eigenvalue of each rotated component, as well as the number of items that loaded on each component in each study. Table 4 presents the same information for the components obtained for the women.

Table 4
Summary of Maximum Between-Study Component Similarity Measures for Women

Component		Component similarity measure	
Reanalysis of Owens' data	Eberhardt & Muchinsky (1982)	Congruence	Correlation
1. Academic achievement (5.49; 13)	2. (5.26; 12)	.92	.91
2. Athletic participation (5.06; 9)	5. (5.13; 9)	.92	.91
3. Social leadership (4.44; 14)	1. (5.62; 19)	.90	.90
4. Parental control (4.34; 11)	3. (4.36; 10)	.92	.92
5. Socioeconomic status (4.32; 10)	4. (4.38; 10)	.89	.89
6. Warmth of paternal relationship (3.86; 9)	10. (3.98; 9)	.87	.86
7. Independence/dominance (3.84; 13)	13. (3.09; 8)	.84	.78
8. Scholastic and cultural interest (3.81; 14)	11. (3.17; 11)	.78	.75
9. Scientific interest (3.47; 11)	12. (5.14; 11)	.88	.88
10. Positive academic attitude (2.93; 9)	14. (3.12; 9)	.88	.88
11. Religious activity (2.90; 4)	6. (3.20; 4)	.84	.84
12. Sibling friction (2.85; 5)	9. (2.51; 3)	.75	.75
13. Warmth of maternal relationship (2.81; 7)	8. (3.32; 11)	.56	.56
14. Negative social adjustment (2.75; 9)	7. (4.17; 16)	.74	.71
15. Approachability (1.85; 4)	No comparable component		

Note. All correlations are significant ($p < .05$). The numbers in parentheses next to each component represent the eigenvalue and the number of items that load on each component, respectively.

It is clear that the components for men that were derived in the reanalysis appear very similar to those obtained by Eberhardt and Muchinsky (1982). The maximum component congruence values range from .67 to .93, with an average of .82. The maximum between-study component correlations for the men range from .64 to .93, with an average correlation of .83. For the women, the components obtained in the reanalysis also appear very similar to those obtained by Eberhardt and Muchinsky. Eliminating Component 15 from consideration, the maximum between-study component congruence values range from .56 to .92, with an average value of .84. The maximum correlations range from .56 to .92, with .84 as the average correlation.

Discussion

Some comment is appropriate relating to differences in the component correlations as reported here, and as had been reported for the Eberhardt and Muchinsky (1982) study. Eberhardt and Muchinsky reported an average maximum correlation of .91 between the components they had obtained for the men in their sample and the components for the men reported in Owens and Schoenfeldt (1979). Two of the component pair correlations were found not to be significant. The comparable value reported here comparing the components obtained for men in the reanalysis with the Eberhardt and Muchinsky components is considerably lower, but then all the correlations were significant. These differences can be accounted for when noting that certain assumptions were made by Eberhardt and Muchinsky about values for the unavailable item loadings from the Owens and Schoenfeldt components. When complete component loading information for the Owens and Schoenfeldt components for men was used in an analysis along with the Eberhardt and Muchinsky loadings, the average maximum between-study component correlation was .83. This value is virtually the same as that obtained when comparing the components obtained for men in the reanalysis with the Eberhardt and Muchinsky components. In effect, the item-loading assumptions used by Eberhardt and Muchinsky had biased the between-study component correlations for men upward in most cases.

The results for the women had been the most troublesome issue of biodata component stability in the Eberhardt and Muchinsky (1982) study. Nearly as many components for

women were not replicated in that study as were. The present results provide strong evidence that the difference in component structures was due to the size of the item pool used, rather than to changing life experiences of women over time. One puzzling issue open for further investigation is to determine why the item responses of women yielded ostensibly different components depending on the additional items involved, whereas this did not occur in the male sample.

It seems clear that the biodata components for both men and women are relatively stable over the time periods and geographical locations for the two samples involved in the present investigation. It should be pointed out that stability is indicated across different samples over time, and not strictly in a longitudinal sense. Biodata components for a college-age population appear to be stable, but the stability of biodata components in a fixed sample over time periods remains an open question. In any event, the stability of biodata components for a college-age population, as evidenced here, should help encourage research into the future utility of biodata as a means for describing, predicting, and understanding behavior.

References

- Burt, C. (1948). The factorial study of temperamental traits. *British Journal of Psychology: Statistics Section*, 1, 178-203.
- Eberhardt, B. J., & Muchinsky, P. M. (1982). An empirical investigation of the factor stability of Owens' biographical questionnaire. *Journal of Applied Psychology*, 67, 138-145.
- Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23, 187-200.
- Owens, W. A., & Schoenfeldt, L. F. (1979). Toward a classification of persons [Monograph]. *Journal of Applied Psychology*, 64, 569-607.
- SAS Institute (1982). *The SAS user's guide (1982 ed.)*. Raleigh, NC: Author.
- Schoenfeldt, L. F. (1974). Utilization of manpower: Development and evaluation of an assessment-classification model for matching individuals with jobs. *Journal of Applied Psychology*, 59, 583-595.
- Tucker, L. R. (1951). *A method for synthesis of factor analysis studies* (Report No. 984). Washington, DC: Department of the Army, Personnel Research Section.
- Wrigley, C., & Neuhaus, J. O. (1955). *The matching of two sets of factors* (Contract Memorandum Rep. No. A-32, Task A). Urbana: University of Illinois.

Received December 20, 1985

Revision received April 21, 1986 ■

Modification of the Minnesota Clerical Test to Predict Performance on Video Display Terminals

Edward M. Silver and Corwin Bennett
Kansas State University

Because of the increased use of video display terminals (VDTs) by clerical workers, the Minnesota Clerical Test (MCT) might need to be modified if it is to predict accurately the performance of clerical workers who use word processors, and new norms might be required for a VDT version of the test. The MCT was administered to 34 subjects in three different forms: paper, VDT, and paper and VDT combined. No alternate forms of the MCT correlated significantly higher than the standard paper and pencil MCT with an interactive word processor task. Also, scores on the MCT administered on a VDT or on a paper and VDT combination failed to explain significantly more variance in performance on the interactive task than did the standard paper MCT alone when these scores were used as predictors in a regression equation. Results suggest that the standard MCT does not require revision; however, the difference in mean performance among the three media suggest that new norms would need to be developed if the MCT were to be administered in whole or in part on a VDT.

The Minnesota Clerical Test (MCT), first published in 1933, has since been revised several times, the latest revision in 1979 (Andrew, Paterson, & Longstaff, 1979). The test is one of the most widely used methods of selecting clerical employees. Although the test is not without flaws (see Kirkpatrick, 1957, and Diaz, 1978), it remains one of the most respected tests of perceptual speed and accuracy. Super and Crites (1962), in a review of clerical aptitude tests, concluded that the MCT had value in clerical selection and that "the higher the score made by a person the higher, other things being equal, he may rise in the field of clerical work" (pp. 177-178).

Because of the increased use of video display terminals (VDTs) by clerical workers, the MCT might need to be modified to predict accurately future performance of word processor operators, and new norms might be required for a VDT version of the test. Projections have been made which suggest that by 1990 "40 to 50 percent of all American workers will have computer terminals" (Cohen, 1983, p. 63). It was suspected that operation of VDTs might require the use of certain eye-hand coordination skills untested by the present form of the MCT.

An example of a clerical job using VDTs is a policy verifier for an insurance company. Policy verifiers check to ensure that information on paper policies corresponds to information displayed on a computer monitor. This type of job is likely to be extremely demanding on perceptual speed and accuracy as well as on skills necessary to operate a video display terminal.

Method

Subjects

The 34 subjects (22 females and 12 males) in this experiment were recruited from the Manhattan, Kansas community. Fifteen subjects re-

ported using a VDT between 5% and 40% of their work time, and 19 subjects reported using a VDT more than 40% of their work time. Most of the subjects were employed as secretaries. The ages of the subjects ranged from 28 to 67 years ($Mdn = 49$ years).

Procedure

Three different forms ("media") of the MCT were used in this experiment: (a) the standard paper and pencil MCT, (b) a version of the MCT entirely administered on a video display terminal (VDT-MCT), and (c) a version of the MCT requiring subjects to compare the left column of names (or numbers) on paper from the original MCT to the corresponding right column on a video display terminal (paper & VDT-MCT). The modified tests were administered in an identical manner to the paper MCT (as described in the instruction manual) with the exception that the test appeared on a VDT (either totally or partially) and that the subjects were required to indicate similarity of columns by depressing keys on a video display terminal rather than by making a pencil check mark between columns. All versions of the MCT consisted of 200 name comparisons and 200 number comparisons. The maximum possible score for each section (names and numbers) for all versions of the MCT was 200. The standard right minus wrong scoring (as described in the MCT manual) was used.

An interactive entry task was also administered to the subjects and used as a criterion. This task required the subjects to type information on a VDT screen in order to fill out an application blank for a fictional organization. Subjects used a directory to supply the names and addresses for the application blank. Subjects were asked to fill out as many application blanks as possible in 30 min. This task was completed by the subjects on two separate occasions. The task was scored by counting the number of pieces of information (names or numbers) correctly entered on the application blank during the allotted time. The mean score for the two separate occasions was used as the interactive task score. This task was included in the study because of its similarity to the type of work performed by word processors and policy verifiers for a local insurance company. The interactive task was therefore assumed to be a reasonable criterion for judging performance of clerical VDT operators.

A repeated measures design was used, with all subjects being administered all forms of the MCT and the interactive task. The order of administration of the various MCT forms was counterbalanced with intervals of approximately 6 weeks between administrations.

This study was conducted prior to Edward Silver's employment with AT&T.

Correspondence concerning this article should be addressed to Edward M. Silver, who is now at AT&T Bell Laboratories (IE 1B-531), P.O. Box 3050, 200 Park Plaza, Naperville, Illinois 60566-7050.

Results

An analysis of variance (ANOVA) of the means shown in Table 1 yielded a statistically significant interaction effect between test media (i.e., paper, VDT, or paper & VDT) and material (i.e., names or numbers), $F(2, 66) = 6.29, p < .01, \omega^2 = .01$. Subjects tended to score especially low in the paper & VDT combined condition when being tested on numeric material. In addition, a significant media effect was found, $F(2, 66) = 95.96, p < .001, \omega^2 = .41$, with subjects tending to score higher on paper tests than on any other media. A significant material effect was also found, with subjects tending to score higher on names than on numbers, $F(1, 33) = 5.86, p < .05, \omega^2 = .01$. (See Table 1.)

Correlations of the name and numeric scores of the three forms of the MCT with the interactive task ranged from .41 to .62 (see Table 1). No modified version of the MCT (names or numeric scores) correlated significantly higher with the interactive task than the paper MCT (names or numeric scores). The test used to determine this nonsignificant difference was Hotelling's T^2 test for correlated coefficients (Guilford, 1956).

Although the paper MCT by itself correlated with the interactive test, as did the modified MCT forms individually, alternate forms of the MCT might be useful in explaining additional variance associated with interactive test scores when used in a multiple regression equation. For example, it is possible that scores on the VDT-MCT might be useful in explaining variance caused by skill in operating a video display terminal. A hierarchical multiple regression equation (see Table 2) was computed to test for this possibility. No significant change in the amount of variance explained (ΔR^2) was found when the VDT-MCT and the paper & VDT-MCT were added to the equation.

Discussion

The first hypothesis was not supported. The results suggest that the standard form of the Minnesota Clerical Test does not need revision to test for clerical aptitude for a task that requires operation of a video display terminal. No alternate medium of the MCT correlated significantly higher with the criterion (the interactive task) than did the standard paper MCT. In addition, the scores of alternate media of the MCT did not significantly add to the amount of variance explained in the criterion by the standard MCT when used as predictors in a regression equation. The second hypothesis was supported. The mean performances for the three media (paper MCT = 243.53; VDT-MCT = 170.18; paper & VDT-MCT = 147.50) indicate the need for restandardization of the test, were it to be administered in whole or in part on a VDT.

Although no significant difference was found between correlations of the standard MCT with the interactive test and correlations of the modified MCTs with the interactive test, the small sample size used in this study ($N = 34$) suggests caution. However, estimations of power in the ANOVA performed in this study indicate that power tended to be high for the Test Material \times Test Media interaction, $\hat{\phi}(2, 66) = 1.88, \text{power} = .78$; for the media effect, $\hat{\phi}(2, 66) = 7.96, \text{power} = .99$; and for the material effect, $\hat{\phi}(1, 33) = 1.56, \text{power} = .50$.

Another limitation of this study was the use of the interactive task as criterion of clerical performance. Although the task was designed to be as similar as possible to the work of policy verifi-

Table 1
Means, Standard Deviations, Correlations, and Reliabilities of Alternate MCT Forms and an Interactive Paper and Video Screen Task

Measure	M	SD	1	2	3	4	5	6	7	8	9	10
Paper MCT												
1. Numeric	119.29	32.42	(.99**) ^a									
2. Names	124.24	33.11	.78**	(.99**) ^a								
3. Combined	243.53	61.81	.94**	.94**	(.99**) ^a							
VDT-MCT												
4. Numeric	85.24	21.15	.64**	.54**	.63**	(.99**) ^a						
5. Names	84.94	19.00	.56**	.79**	.72**	.54**	(.96**) ^a					
6. Combined	170.18	35.26	.69**	.75**	.76**	.89**	.86**	(.98**) ^a				
PAPER & VDT-MCT												
7. Numeric	67.32	15.98	.32	.45**	.41*	.36*	.52**	.50**	(.97**) ^a			
8. Names	80.18	20.42	.41*	.61**	.55**	.42*	.75**	.66**	.70**	(.97**) ^a		
9. Combined	147.50	33.56	.41*	.59**	.53**	.43*	.70**	.64**	.90**	.94**	(.99**) ^a	
10. Interactive task	761.96	205.38	.48**	.56**	.55**	.56**	.52**	.62**	.41*	.52**	.51**	(.86**) ^b

Note. $N = 34$. MCT = Minnesota Clerical Test; VDT = video display terminal.
* $p < .05$. ** $p < .01$.
^a Split-half (odd-even) reliability estimate adjusted by the Spearman-Brown prophesy formula.
^b Test-Retest reliability estimate.

Table 2
Results of Multiple Regression Predicting
Interactive Test Scores

Independent variable	Step 1	Step 2	Step 3
Paper MCT			
Numeric			
Names			
R^2	.32**		
R^2 adjusted	.27		
VDT-MCT			
Numeric			
Names			
R^2		.41**	
ΔR^2		.10	
R^2 adjusted		.33	
Paper & VDT-MCT			
Numeric			
Names			
R^2			.45**
ΔR^2			.03
R^2 adjusted			.32

Note. $N = 34$. MCT = Minnesota Clerical Test; VDT = video display terminal.

** $p < .01$.

ers, there were some differences. The interactive task was of short duration (two ½-hr sessions), and the subjects did not have to worry about the possible negative consequences of poor performance that would likely occur in a job.

Despite these limitations, these results suggest caution to developers of revised versions of standardized tests. According to Standard 4.6 of the 1985 edition of the *Standards for Educational and Psychological Tests*, scores on different forms of a test, including computerized adaptive tests, must include data concerning the parallelism of the tests. The modified versions

of the MCT were developed in an attempt to improve the original MCT's predictive accuracy, and were not intended to be an equivalent form of the MCT. Nevertheless, the findings in this study indicate that even minor modifications of testing procedures can have significant impact on testing norms. According to Standard 4.7 of the 1985 edition of the *Standards for Educational and Psychological Tests*

When radical shifts in test specifications occur, either a new scale should be introduced or a clear statement should be provided with the scores to alert users that the scores are not interchangeable with those on earlier versions of the test. (p. 34)

Results of the present study suggest that shifts in test specification do not have to be radical to produce major changes in scoring norms. Modified test versions should be accompanied by revised norms and expectancy tables.

References

- Andrew, D. M., Paterson, D. G., & Longstaff, H. P. (1979). *Manual for the Minnesota Clerical Test*. New York: Psychological Corporation.
- Cohen, M. M. (1983). Computer prescribing: Help patients combat 'VDT' fatigue. *Review of Optometry*, 120, 55-63.
- Diaz, A. P. D. L. (1978). Construction defect in the Minnesota Clerical Test. *Professional Psychology*, 9, 7-8.
- Guilford, J. P. (1956). *Fundamental statistics in psychology and education*, (4th ed.). New York: McGraw-Hill.
- Kirkpatrick, D. L. (1957). The Minnesota Clerical Test. *Personnel Psychology*, 10, 53-54.
- Standards for Educational and Psychological Testing*. (1985). Washington, DC: American Psychological Association.
- Super, D. E., & Crites, J. O. (1962). *Appraising vocational fitness*. New York: Harper & Brothers.

Received May 27, 1986

Revision received September 8, 1986 ■

Sex Effects in Workplace Justice Outcomes: A Field Assessment

Dan R. Dalton

Graduate School of Business, Indiana University

William D. Todor

College of Administrative Science, Ohio State University

Crystal L. Owen

College of Administrative Science, Ohio State University

Recent research on sex effects in workplace justice leads to different conclusions based largely on whether simulations or field assessments were relied on for experimental protocol. Our examination provides a field assessment that replicates in part and extends the investigation of dispute and disciplinary outcomes in the workplace. We argue that it is not merely the sex of the actor, but the nature of the dyadic composition (man/man, woman/woman, man/woman, woman/man) that is a critical influence on outcomes. This hypothesis is supported by significant differences in workplace justice outcomes based on the dyadic composition of the actors involved in dispute proceedings ($N = 498$).

Although there is a modest literature on workplace justice that considers dispute or disciplinary proceedings in the workplace, it leads to a uniform, troubling conclusion: Systematic differences in outcomes associated with workplace justice may be related to the sex of the involved employee (e.g., Dalton & Todor, 1985a; Dobbins, 1985; Dobbins, Pence, Orban, & Sgro, 1983; Larwood, Rand, & Hovanessian, 1979; Rosen & Jerdee, 1975). The nature of these differences is difficult to summarize, however, because this research reports inconsistently with respect to the direction of bias, that is, whether women or men are more likely to prevail. We seek here to replicate in part and extend this work on sex effects in workplace justice in a field assessment.

Comparison of Simulations With Field Assessments

Research relying on simulation (Dobbins, 1985; Dobbins et al., 1983; Larwood et al., 1979; Rosen & Jerdee, 1975) has reached consistent conclusions: Women may be treated quite differently and less favorably both in the adjudication of disputes and in disciplinary proceedings. Larwood et al., for instance, reported that disciplinary actions would be more readily taken against a woman.

Three recent field assessments examined sex factors in workplace justice and reached different conclusions. Dalton and Todor (1985b) found no evidence of differences in workplace justice outcomes for women and men. Beyer and Trice (1984) reported that women were less likely to be suspended or discharged than their male counterparts. Another examination concluded that women in the workplace were more likely than men to prevail in dispute and disciplinary hearings (Dalton & Todor, 1985a).

That results of discrimination research under laboratory protocol are inconsistent with those in field settings has been noted

(e.g., Dalton & Todor, 1985a; Osborn & Vicars, 1976; Wendelken & Inn, 1981). Osborn and Vicars speculated on such differences: "Artificial, short-term laboratory situations tend to elicit subject responses based on readily available stereotypes, while long-term, real life, field settings include extensive interpersonal contact that provides subjects with a more realistic basis for their behavior" (1976, p. 447).

Such tendencies may also be consistent with the notion of cognitive simplicity or complexity (e.g., Bieri, 1955; Schroder, Driver, & Streufert, 1967). Here again, it might be argued that the typical laboratory subject may be relatively unversed in workplace justice proceedings and may respond somewhat stereotypically. Individuals with field experience, on the other hand, may be privy to more information. It may be that these factors lead to different outcomes. Also, simulation subjects are not accountable for their decisions.

Sex Effects Versus Sex Context

The dyadic composition of the actors in workplace disputes or disciplinary actions in the workplace may also be an issue. For most organizations, dispute and disciplinary proceedings are dyadic in nature: An employee may have some dispute with the supervisor; an employee may be subject to disciplinary action by the supervisor. To examine only the sex of the employee to determine differences in workplace justice outcomes, then, may be insufficient (Dalton & Todor, 1985b; Dobbins, 1985; Dobbins et al., 1983).

Dobbins (1985), for example, reported a variety of interactions of leader sex with subordinate sex in the manner by which leaders respond to poor performance by subordinates. Dalton and Todor (1985b), in a field assessment, have also reported substantial differences in certain workplace outcomes as a function of the sex composition of the actors.

A key to understanding apparent differences in workplace justice outcomes, then, may be the sex composition and nature of the interaction between complainants and those to whom complaints are brought.

Correspondence concerning this article should be addressed to Dan R. Dalton, Department of Management, Graduate School of Business, Indiana University, Bloomington, Indiana 47405.

Hypotheses

To assess the effects of sex on workplace justice outcomes, it is necessary to identify some organizational process to which issues of discipline and dispute are regularly subjected. For these purposes, the grievance process seems especially well suited (Dalton & Todor, 1985a). In addition, it is a unique process in that it allows the sex of the involved employee and the dyadic composition of those with the mission to adjudicate the problem to be separated.

Obviously, the sex of involved employees can be ascertained and its relation with dispute and disciplinary outcomes can be easily determined. Interestingly, the involved employee is not a party to the dyad that will interact to resolve the dispute. This is the mission of a dyad comprised of a union representative and a company supervisor (e.g., Dalton & Todor, 1982). The sex of the involved employee and the effects of the dyadic composition of the actual actors to the resolution of the dispute can be examined independently.

Hypothesis 1: There will be a systematic difference in workplace justice outcomes (the proportion of outcomes in which complainants prevail) based on the sex of the involved employee.

Hypothesis 2: There will be a systematic difference in workplace justice outcomes (the proportion of outcomes in which complainants prevail) based on the dyadic sex composition of the actors (union representative and company supervisor) in dispute and disciplinary proceedings.

It has been demonstrated that the outcomes of grievances are in part a function of their viability and severity (Dalton & Todor, 1985a). The effects of these factors, then, will be considered in subsequent analyses. This viability variable recognizes that all grievances do not have the same a priori probability of winning (e.g., Dalton & Todor, 1981, 1985a; Fossum, 1985). It may be that female-supervisor/male-union-representative dyads, for example, handle less viable complaints—those with a lesser a priori probability of favorable outcome. Observed differences, then, would not be based on sex or dyadic effects, but predicated on the review of less viable cases.

The severity of the issue antecedent to the dispute or disciplinary proceeding is a possible confound in this research as well. Although it may be true that one sex or dyadic composition is associated with different outcomes, there may be systematic differences in the nature of disputes and disciplinary proceedings that would affect those outcomes. It would not be surprising to find that less serious matters are more likely to meet a favorable outcome.

Method

Sample

The sample for this examination is all grievances ($N = 673$) filed by employees over a 1-year period in a large labor union local that represents a public utility. The data for this report were derived from archival records.

Variables

Sex. For each grievance, the sex of the involved employee was determined.

Dyadic sex composition. The sex of the union representative and the supervisor assigned to the dispute were determined and assigned to the four possible dyadic configurations: male supervisor with male union representative, male supervisor with female union representative, female supervisor with female union representative, and female supervisor with male representative. These configurations, then, constitute the sex composition variable.

Grievance resolution. Grievance resolution is subject to five outcomes: win, lose, compromise, withdraw without prejudice, and abandon. For practical purposes, several of the resolution categories are the same. Although there are technical (and legal) differences among lost, withdrawn, and abandoned actions, the outcome is the same from the employees' view—their demands have not been met. These categories can be combined for analysis, inasmuch as they all represent unfavorable outcomes for the employee (Dalton & Todor, 1981, 1985a).

The compromise category may present a problem. To determine whether a given compromise represents a favorable or unfavorable outcome for an employee would be hazardous. Is a 5-day suspension reduced to a 3-day suspension favorable or unfavorable? Grievances in which a compromise was reached, therefore, are excluded from the analysis. Although this is not a critical exclusion, inasmuch as the proportions of compromise outcomes are not significantly different across sex of the complainant or among the dyadic compositions, it does reduce the sample size to 498.

For analytical purposes, then, there are two outcome categories: An individual can win—a favorable outcome; the individual can lose, the complaint is withdrawn, or it is abandoned—an unfavorable outcome.

Grievance viability. Grievance viability is not subject to direct assessment. A reasonable surrogate can be determined, however, by noting the number of levels to which a grievance has proceeded before resolution (Dalton & Todor, 1981, 1985a). For this sample, there are five levels through which a grievance could proceed: the local, division, company, national, and arbitration. Analysis includes only the first four levels, inasmuch as no cases proceeded to arbitration for this sample.

Grievance severity. Grievance severity is easily coded to establish two distinct types. The first category can be described as *serious*, inasmuch as actions here are disciplinary and job threatening (e.g., suspension, termination, disciplinary memoranda). The second category is neither job threatening nor disciplinary in nature (Dalton & Todor, 1981, 1985a). Grievances in this category would include overtime by-pass, seniority issues for choice of shift, and vacation preference.

Results

The concern of this research is the difference, if any, in workplace justice outcomes as a function of the sex of the involved employee or the dyadic composition of those actors assigned to resolving these disputes.

Table 1 is a descriptive illustration of the proportion of favorable outcomes for grievants as the outcomes are related to sex composition, grievance severity, and viability (level of resolution). This table also includes descriptive data on the percentage of male and female supervisors and union representatives for each sample.

There are substantial differences in the "winning percentage" for employees depending on dyadic composition. Comparing the extreme cases (female-supervisors/male-union-representatives vs. male-supervisor/male-union-representative) indicates that the latter dyad is more than three times as likely to support the employee.

It appears, then, that there are gross differences in workplace justice outcomes as a function of the dyadic composition of the

Table 1
Proportions of Workplace Justice Outcomes, Grievance Severity, and Grievance Viability as a Function of Sex Composition

Dyadic composition	Favorable outcomes/ unfavorable outcomes	Serious grievances/ nonserious grievances	Resolved 1st level/ resolved > 1st level
Male supervisor with male union representative	40.5/59.5	23.7/76.3	56.2/43.8
Female supervisor with female union representative	22.6/77.4	28.2/71.8	46.3/53.7
Female supervisor with male union representative	13.2/86.8	28.1/71.9	47.2/52.8
Male supervisor with female union representative	23.6/76.4	20.1/79.9	58.0/42.0

Note. 62.6% of the supervisors were men and 37.4% were women; 53.3% of the union representatives were men and 46.7% were women.

actors. Given the categorical nature of the variables, we rely on a loglinear analysis to test our hypotheses.

As previously noted, both viability and severity of disputes have been shown to be robust predictors of grievances outcomes (Dalton & Todor, 1985a). The issue here is whether a logit model including employee sex and/or the dyadic composition variable is a better model. Our initial model specification, including only viability and severity, indicates a likelihood ratio chi-square of 140.37 ($p < .001$). Adding employee sex to the logit model results in a likelihood ratio chi-square of 138.72 ($p < .001$). Given the differences in degrees of freedom for the models, the slight improvement for this model is not statistically significant. It can be concluded, therefore, that employee sex does not contribute to differences in grievance outcomes.

Adding the dyadic composition variable to the model results in an improved likelihood ratio chi-square of 122.58 ($p < .001$). Subtracting the likelihood ratios to determine the improvement in this model, we obtain a chi-square of 17.79 with four degrees of freedom ($p < .01$). It can be concluded, therefore, that the dyadic composition variable does add significantly to the model; that is, the dyadic composition of the actors does have a significant effect in determining grievance outcomes. Examination of the parameter estimates indicates no contribution from interactive effects.

Similar results can also be demonstrated by relying on alternative analyses that yield effect sizes for these variables as well. A hierarchical regression can be used with outcome as the dependent variable with viability and severity entered first, sex of the complainant, and the sex composition (recoded same-sex/cross-sex) variable entered subsequently. This analysis indicates a total explanatory power of just over 17% variance explained. The sex composition accounts for some 3.3% unique variance of that total. Complainant sex provides no significant contribution.

Discussion

These results have potentially serious implications. It is reassuring that sex of the complainant does not seem to have an impact on workplace outcomes. The suggestion that employee outcomes are systematically related to the dyadic composition of those individuals charged with the responsibility of “hearing their cases” is disturbing, particularly because similar reports have been noted (e.g., Dalton & Todor, 1985b; Dobbins, 1985; Dobbins et al., 1983).

Although we report only a phenomenon, there are patterns of related research that may in part account for such effects. Zammuto, London, and Rowland (1979), for example, reported that the sexual composition of supervisor/employee dyads is an important predictor of the nature of conflict resolution strategies. They concluded, for instance, that male/male interactions were more likely to result in a confrontational strategy than female/female interactions. It is notable that male/male confrontations are more likely to result in employee “wins” than are female/female confrontations for this sample. Perhaps the reliance on confrontational strategy between union and company representatives does result in better outcomes for aggrieved employees.

Also, recent research has demonstrated that reward allocations may vary, not only by the sex of the allocator, but also by the dyadic composition of the allocator/recipient (e.g., Kahn, Nelson, & Gaeddert, 1980; Reis & Jackson, 1981). Reis and Jackson, for example, reported that both sexes allocated more generously to an opposite-sex partner than to a same-sex partner, irrespective of task type or performance. It may be that actors in certain dyadic compositions are likely to make concessions in workplace justice proceedings, and these in turn, lead to different outcomes.

Dobbins et al. (1983) may also provide an interesting explanation. They have reported significant differences in the tendency for same-sex/cross-sex dyads to attribute causation. Subjects in their same-sex dyads made greater external attributions for described phenomenon than cross-sex dyads. It may be that the perceived reason (e.g., internal/external attribution) for a given incident or condition that led to a grievance is an important factor in its resolution. If sex composition dyads differ in their attributional biases, different outcomes in the grievance proceedings may result.

There has been little field research conducted in the area of workplace justice processes. Given the apparent ambiguity between field and simulation reports, and the practical—and legal—implications of the results, perhaps replication beyond these data and the generation of related research concerning workplace justice would be fruitful areas of investigation.

References

Beyer, J. M., & Trice, H. M. (1984). A field study of the use and perceived effects of discipline in controlling work performance. *Academy of Management Journal*, 27, 743–764.

- Bieri, J. (1955). Cognitive complexity-simplicity and predictive behavior. *Journal of Abnormal and Social Psychology*, 51, 263-268.
- Dalton, D. R., & Todor, W. D. (1981). Win, lose, draw: The grievance process in practice. *Personnel Administrator*, 6, 54-59.
- Dalton, D. R., & Todor, W. D. (1982). Antecedents of grievance filing behavior: Attitude/behavioral consistency and the union steward. *Academy of Management Journal*, 25, 158-169.
- Dalton, D. R., & Todor, W. D. (1985a). Gender and workplace justice: A field assessment. *Personnel Psychology*, 38, 133-151.
- Dalton, D. R., & Todor, W. D. (1985b). Composition of dyads as a factor in the outcomes of workplace justice. *Academy of Management Journal*, 28, 704-712.
- Dobbins, G. H. (1985). Effects of gender on leaders' responses to poor performers: An attributional interpretation. *Academy of Management Journal*, 28, 587-598.
- Dobbins, G. H., Pence, E. C., Orban, J. A., & Sgro, J. A. (1983). The effects of sex of the leader and sex of the subordinate on the use of organizational control policy. *Organizational Behavior and Human Performance*, 32, 325-343.
- Fossum, J. A. (1985). *Labor relations*. Plano, TX: Business Publications.
- Kahn, A., Nelson, R. E., & Gaeddert, W. P. (1980). Sex of the subject and sex composition of the group as determinants of reward allocations. *Journal of Personality and Social Psychology*, 38, 737-750.
- Larwood, L., Rand, P., & Hovanessian, D. (1979). Sex differences in response to simulated employee discipline cases. *Personnel Psychology*, 32, 539-550.
- Osborn, R. N., & Vicars, W. M. (1976). Sex stereotypes: An artifact in leader behavior and subordinate satisfaction analysis. *Academy of Management Journal*, 19, 439-449.
- Reis, H. T., & Jackson, L. A. (1981). Sex differences in reward allocation: Subjects, partners, and tasks. *Journal of Personality and Social Psychology*, 40, 465-478.
- Rosen, B., & Jerdee, T. H. (1975). Effects of employee's sex and threatening versus pleading appeals on managerial evaluations of grievances. *Journal of Applied Psychology*, 60, 442-445.
- Schroder, H. M., Driver, M. J., & Streufert, S. (1967). *Human information processing: Individuals and groups functioning in complex social situations*. New York: Holt, Rinehart, & Winston.
- Wendelkin, D. J., & Inn, A. (1981). Nonperformance influences of performance evaluations: A laboratory phenomenon? *Journal of Applied Psychology*, 66, 149-158.
- Zammuto, R. F., London, M., Rowland, K. M. (1979). Effects of sex on commitment and conflict resolution. *Journal of Applied Psychology*, 64, 227-231.

Received January 3, 1986

Revision received May 1, 1986 ■

Predictive Validity of the MODE Conflict Instrument

Boris Kabanoff

Australian Graduate School of Management
University of New South Wales, New South Wales, Australia

Sixty-three members of an MBA (Master of Business Administration) class completed the MODE instrument, which measures preferences for five conflict modes (competing, collaborating, avoiding, compromising, and accommodating), in addition to a number of personality measures. One year later, people were rated on their observed use of these five conflict modes over the year. Little association was found between MODE scores and rated conflict behavior, but several of the personality measures were correlated with the ratings.

Blake and Mouton (1964) made a significant contribution to the description of interpersonal conflict behavior by introducing a five-category scheme for classifying interpersonal conflict-handling modes. Their scheme includes five modes—competing, collaborating, compromising, avoiding, and accommodating. Later, Thomas (1976) interpreted these five modes as reflections of two underlying cognitive/affective dimensions: *cooperation* (attempting to satisfy the other person's concerns) and *assertiveness* (attempting to satisfy one's own concerns). Competing is assertive and uncooperative; collaborating is cooperative and assertive; avoiding is unassertive and uncooperative; accommodating is cooperative and unassertive; and compromising is intermediate in both cooperativeness and assertiveness.

A number of instruments have been developed to measure people's preferences for these five conflict-resolution styles (Blake & Mouton, 1964; Hall, 1969; Lawrence & Lorsch, 1967). However, research by Thomas and Kilmann (1973, 1975) indicated that responses to these measures were overwhelmingly influenced by the social desirability of the conflict-handling modes and their phrasings. In response to this, Kilmann and Thomas (1977) developed a new measuring instrument—the Management-of-Differences Exercise, or MODE instrument.

MODE is an ipsative questionnaire consisting of 30 sets of paired items, with each item describing one of the five conflict modes. Each mode is paired with each other mode three times, and a person's score on each mode (range = 0–12) is the number of times statements representing that mode are selected over other statements. Kilmann and Thomas (1977) provided evidence for a decreased level of social desirability, acceptable lev-

els of internal and test-retest reliability, and some initial evidence of convergent validity for MODE.

However, as Kilmann and Thomas (1977, p. 319) observed—and their observation is still current—the MODE instrument has been used in only a few research settings and we do not yet have the kind of results that would give us strong evidence for predictive validity, that is, whether MODE scores predict actual conflict behavior. Yet evidence of predictive validity is generally the most rigorous and demanding test of the usefulness of an instrument in both research and applied contexts. The present study assesses the predictive validity of the MODE instrument by examining the association between MODE scores and people's observed conflict behavior over an extended period. It also compares the predictive validity of the MODE instrument with a number of established personality measures.

Method

Participants

Participants were students ($N = 78$) in a full-time (MBA) course at an Australian university. Their mean age was 27.6 years; and 87% were men. Most of the participants had worked for at least five years in administrative, professional, or managerial occupations. The majority were native Australians, and approximately 30% came from overseas, primarily Southeast Asia.

Procedure

During a first-year organizational-behavior course, most students ($n = 63$) completed the MODE instrument and a majority ($n = 46$ –57) also completed the personality measures described later in this article. Completion of the scales was voluntary, although students were told that they would receive group and individual feedback during a future class.

Twelve to 15 months after this information was gathered, all of the class members were sent a letter and a questionnaire seeking their cooperation in this study. Students were asked to rate their peers on their observed use of one of the five conflict styles. Because class members had had many opportunities to interact in a variety of work and social settings, it seemed reasonable to assume that they had formed some fairly clear impressions of how peers dealt with conflict situations.

Ratings were done using a questionnaire that contained the names of all of the class members and five columns that described the likelihood of a person using a conflict mode. The response categories were “fairly

This research was made possible by grants from the University of New South Wales, the Australian Graduate School of Management, and the Australian Research Grants Scheme.

I am grateful to Xicom Corporation for granting permission to use the MODE instrument, and also to the participating members of my MBA class.

Correspondence concerning this article should be addressed to Boris Kabanoff, Australian Graduate School of Management, University of New South Wales, P.O. Box 1, Kensington, New South Wales, 2033, Australia.

likely," "somewhat likely," "neither likely nor unlikely," "somewhat unlikely," and "fairly unlikely." There was also a "don't know" option. The use of global ratings had a practical advantage in the present study inasmuch as participants rated a large number of people; however, global ratings also increased the potential for biases such as halo. Nevertheless, this was considered a necessary trade-off.

Raters read a description of each mode, based largely on Thomas and Kilmann (1974). The instructions stressed the need to attend to the behavioral characteristics of the style described. On the basis of previous experience with the actual MODE titles in an interpersonal rating context, I decided to use different, more neutral titles for each of the styles that follow.

Let's do it my way (i.e., competing). Let's do it my way means that a person stands up for his or her rights, defends a position that he or she believes in, or simply tries to get his or her views adopted. This style means that a person's primary emphasis is on ensuring that ideas and views that he or she believes in are accepted by the other party. It reflects a commitment to a point of view and an attempt to convince the other person(s) of the merits of a position. This is a highly assertive style.

Split the difference (i.e., compromising). This mode, split the difference, involves giving the other party some of the things that he or she wants in return for getting some of the things you want. Neither party has all of his or her needs and aims met, but both parties get something. This mode reflects a willingness to trade or bargain with the other party and to search for a middle ground. This is a pragmatic mode that involves negotiation and searching for a relatively quick solution to the problem.

Better let the situation cool down before we act (i.e., avoiding). This mode involves not doing anything directly about the issue when it arises, but trying a cooling down period in which two or more people who are in conflict can postpone the argument until everyone is less likely to lose their tempers. One could describe this mode as diplomatically putting the issue to one side so that one avoids the immediate onset of conflict and waits until there is a better time for people to try and handle it. This mode essentially puts resolution of the issue into the future, with the view that the passage of time may in itself solve the problem, or that people will be more agreeable with one another if they avoid a fight right at the beginning. This mode may involve withdrawing from the conflict, agreeing to put it aside, or successfully side stepping the issue.

Maybe we can work this one out (i.e., collaborating). This mode involves two or more people confronting an issue and trying to resolve it in a mutually satisfactory way. Each person will be trying to achieve his or her own goals, and for this approach to succeed, the parties involved should be able to find a resolution to the issue that meets both their desires. This mode requires exploring a disagreement in depth and requires considerable sharing of views and information in an atmosphere that encourages the other person to be frank and open, by being frank and open yourself. It is a mode that requires creativity from both parties in order to be successful.

I see your point of view (i.e., accommodating). Following the accommodating mode, a person is willing to agree with the other person because he or she sees the other person's point of view or perhaps because he or she sees some short-term or long-term benefit to agreeing with the other person, such as preserving good relations for the future. This mode may involve attempts to soothe the other person's feelings in order to stop the further development of conflict, to appear agreeable to the other person, and to be seen as reasonable and friendly. This mode implies a strong concern for the views of the other person and a somewhat lesser concern with one's own desires.

Class members also completed four personality measures—Machiavellianism (Christie & Geis, 1977), locus of control (Rotter, 1966), and expressed needs for control and inclusion (Schutz, 1958). Locus of control and Machiavellianism are probably the most frequently used per-

Table 1
Statistics for Major Variables

Variable	<i>M</i>	<i>SD</i>	<i>N</i>
COMPETING	6.19	2.93	63
COLLABORATING	6.59	2.43	62
COMPROMISING	5.92	2.59	63
AVOIDING	6.11	2.53	62
ACCOMMODATING	4.69	2.39	63
Machiavellianism	37.87	12.04	56
Locus of control	7.03	3.46	57
Control need	3.72	2.48	46
Inclusion need	4.52	2.14	46
Competing	3.05	0.81	78
Collaborating	3.62	0.53	76
Compromising	3.32	0.49	74
Avoiding	3.53	0.60	76
Accommodating	3.22	0.56	77

Note. MODE scores are capitalized.

sonality measures in organizational research. The Machiavellianism scale has previously been used in research involving MODE (Kilmann & Thomas 1977). Expressed needs for control and inclusion were measured as potential personality influences on assertiveness and cooperativeness, respectively, which are posited to underlie these five conflict styles. It was hypothesized that persons with a higher control need would be more likely to use assertive conflict modes, especially *competing*, whereas *inclusion*, which Schutz (1958) defined as a high need to establish and maintain satisfactory relationships, would be positively correlated with people's use of cooperative modes such as collaboration, compromise, and accommodation.

The personality measures served two purposes. First, there was an inherent interest in the correlations between personality and both the self-reported preference for and the rated use of the different conflict modes. A second role for the personality measures was as benchmarks against which the predictive validity of the MODE instrument could be compared.

Results

Of the 77 distributed questionnaires, 54 were returned after one reminder. The modal number of raters for each mode, excluding those choosing the "don't know" option, was as follows: competing (9), collaborating (7), compromising (9), avoiding (7), and accommodating (9). Excluding cases with fewer than 3 raters resulted in the loss of between 1 and 3 cases for each mode.

Interrater reliability was calculated directly from the variance-covariance matrix (Winer, 1971, pp. 291–295), which was derived using all available pairs of cases, excluding those just described. Two raters for the compromising mode were excluded owing to their low covariance with the other raters. Using the Spearman-Brown correction formula, interrater reliabilities for the five modes were as follows: competing (.82), collaborating (.50), compromising (.59), avoiding (.72), and accommodating (.78). Overall, these levels of reliability were considered satisfactory; therefore, people's observed use of the conflict modes was calculated as their average score over all available raters, with higher scores indicating higher rated likelihood of using the mode. Table 1 shows means, standard deviation, and sample sizes for all variables used in the subsequent analyses.

Table 2
Correlation Between Behavior Ratings

Variable	1	2	3	4	5
Behavior ratings					
1. Competing	—				
2. Collaborating	-.31*	—			
3. Compromising	-.61*	.66*	—		
4. Avoiding	-.49*	.65*	.64*	—	
5. Accommodating	-.67*	.54*	.74*	.61*	—

* $p < .01$.

Table 2 shows the Pearson product-moment correlations between people's rated scores on the five conflict modes. It is evident that people's ratings on the five modes are significantly correlated despite being gathered from independent judges. The correlations suggest that people's conflict behavior was judged on a general competing-noncompeting dimension. However, all five scores were retained for further analysis inasmuch as a partial-correlation procedure could be used.

The correlations in Table 3 are between rated scores and people's MODE scores. There are no significant correlations between MODE scores and mode ratings. All of the critical correlations between corresponding MODE and rated scores are close to zero, and three are negative. Fourth-order partial correlations controlling for correlations between the behavior ratings did not alter the general picture though several of the values changed. Partial correlation produced a significant positive association between accommodating score on MODE and rated accommodating, but also revealed a significant negative association between MODE avoiding and rated avoiding. Thus, overall, we may conclude that there is little agreement between MODE score and behavior ratings. Given this, the correlations in Table 4 between personality and both sets of conflict scores are particularly interesting.

The main pattern of correlations is between control need and both MODE scores and behavior ratings. As expected, people with a stronger control need were more likely to choose a *competing* MODE and less likely to choose an *avoidance* MODE. However, control was uncorrelated with preference for *accommodating*. Turning to the fifth-order partial correlations (controlling for the correlations between behavior ratings and control and inclusion, $r = .45$, $p < .01$), it was found that control

need was positively correlated with rated competing and negatively correlated with compromising, consistent with expectations. Contrary to expectations, control need tended to be positively correlated with rated accommodating. Although this unexpected result requires replication, it is possible that people with a win or lose orientation to interpersonal conflict (implied by a high control need) have a flip-flop conflict pattern because of their nonuse of the intermediate, compromising style.

Partial correlation, controlling for the correlation between the two need measures revealed that the apparent positive correlation between inclusion need and preference for competing was spurious. Inclusion was significantly negatively correlated with rated competing, but was not positively correlated with the more cooperative modes such as collaborating, compromising, and accommodating, as had been predicted. Neither Machiavellianism nor locus of control were associated with any of the conflict measures.

Discussion

Despite limitations in the design and methodology adopted by this study, these results cast doubts on the predictive validity of the MODE instrument. Thus, a continuing role for the MODE questionnaire as a behavior-diagnostic device may be unwarranted. However, there are several caveats to this conclusion.

The behavior ratings obtained in this study were intercorrelated and suggest the presence of a significant halo effect on a general competing-noncompeting dimension. This threatens the validity of the study to a degree. However, despite this halo effect one would have expected to find, if MODE scores are valid predictors, that the two styles that fall at opposite ends of this continuum (i.e., competing and accommodating) were correlated with their respective MODE scores. This was true only for accommodating, whereas competing, the most reliably rated style, displayed no association between MODE and rated scores.

A more general and perhaps more serious threat to this study is that there are no real, behavioral equivalents of these conflict styles that can be identified independently of the context in which they occur. That is, behavior that is competing in one context is not perceived as competing in another, so that the notion of measuring general styles of conflict behavior is inherently invalid. This, of course, is an objection not only to this study but, a priori, to the validity of the MODE instrument, and has been directed at many other "behavior-diagnostic" instruments such as personality tests (e.g., see Jackson & Pauno-

Table 3
Correlations Between Behavior Ratings and MODE Scores

Variable	MODE scores				
	Compete	Collaborate	Compromise	Avoid	Accommodate
Competing	-.04 (.05)	.05	.01	-.13	.06
Collaborating	-.18	.05 (-.01)	.08	.14	.11
Compromising	-.07	.22*	-.02 (-.07)	-.04	.03
Avoiding	-.04	-.03	-.04	-.04 (-.21)*	-.01
Accommodating	.04	-.12	.05	-.08	-.17 (.27)**

Note. Numbers in parentheses are fourth-order partials controlling for ratings.
* $p < .10$. ** $p < .05$.

Table 4
Correlation of MODE Scores and Ratings With Personality

Variable	Personality measures			
	Machiavellian	Locus of control	Control	Inclusion
MODE scores				
COMPETE	.12	.07	.49***	.30** (.10) ^a
COLLABORATE	-.12	.19	.05	.03
COMPROMISE	-.04	-.03	.04	.13
AVOID	.20	.20	-.28*	.07
ACCOMMODATE	-.18	.02	.06	.00
Behavior ratings				
Competing	-.17 (-.19) ^b	.05 (-.04) ^b	.40*** (.37)** ^c	-.06 (-.28)** ^c
Collaborating	.02 (.09)	.02 (.07)	-.34** (-.18)	.05 (.12)
Compromising	-.06 (-.13)	-.06 (.00)	-.45*** (-.27)**	-.05 (-.01)
Avoiding	.06 (.16)	.04 (-.09)	-.17 (.04)	.20 (.15)
Accommodating	-.04 (-.10)	-.08 (-.07)	-.28* (.26)*	.01 (-.15)

Note. Numbers in parentheses are ^a first-order partial controlling for control, ^b fourth-order partials controlling for ratings, and ^c fifth-order partials controlling for ratings and control or inclusion.

* $p < .10$. ** $p < .05$. *** $p < .01$.

nen's, 1980, review). Only a limited answer to this criticism can be offered here. The level of interrater agreement obtained in this study suggests (a) that people were fairly consistent in the way they described different people's behavior and (b) that they did not find the task of extracting behavioral meaning from its social context an impossible one. Furthermore, at least some of the personality measures that are open to the same objection did display a regular pattern of association with behavior ratings, indicating that they have a predictive validity not possessed by the MODE measure. Nevertheless, this is a significant and fundamental challenge to MODE and to similar instruments, and points to the need for full-blown validation studies rather than mere refinements of wording and format.

Can MODE play any useful role in applied and research contexts? It could be argued that in applied settings MODE may help people learn about their conflict intentions even if it did not describe their actual behavior. The difficulty with this suggestion is that MODE is presented as a behavior-diagnostic device and, in any case, that most people may have difficulty discriminating between intentions and behavior. However, this suggests an interesting research issue. Jackson and Paunonen (1980) noted that people seem to differ significantly in their level of behavioral self-knowledge. Hence, an interesting and potentially useful question is what characteristics facilitate or inhibit people's accurate reporting and perhaps their awareness of their own behavior in conflict situations? Other issues that may be addressed by future studies include whether these results hold for other groups of respondents, the effects of feedback on people's future MODE preferences and means for improving the behavioral ratings. However, the major conclusion to be drawn from this study is that, subject to further research, there are reasons to doubt the utility of the MODE instrument as a means of predicting conflict behavior.

References

- Blake, R. R., & Mouton, J. S. (1964). *The managerial grid*. Houston, TX: Gulf Publishing.
- Christie, R., & Geis, F. L. (1970). *Studies in Machiavellianism*. New York: Academic Press.
- Hall, J. (1969). *Conflict management survey: A survey on one's characteristic reaction to and handling of conflicts between himself and others*. Conroe, TX: Teleometrics International.
- Jackson, D. N., & Paunonen, S. V. (1980). Personality structure and assessment. *Annual Review of Psychology*, 31, 503-551.
- Kilmann, R. H., & Thomas, K. W. (1977). Developing a forced-choice measure of conflict handling behavior: The "Mode" instrument. *Educational and Psychological Measurement*, 37, 309-325.
- Lawrence, P. R., & Lorsch, J. W. (1967). *Organization and environment*. Boston: Harvard University, Graduate School of Business Administration.
- Rotter, J. B. (1966). Generalized expectancies for internal versus external control. *Psychological Monographs*, 80 (1, Whole No. 609).
- Schutz, W. C. (1958). *FIRO: A three dimensional theory of interpersonal behavior*. New York: Rinehart.
- Thomas, K. W. (1976). Conflict and conflict management. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 889-935) Chicago: Rand-McNally.
- Thomas, K. W., & Kilmann, R. H. (1973). *Some properties of existing conflict behavior instruments* (Working paper No. 72-11). Los Angeles: University of California Graduate School of Management, Human Systems Development Center.
- Thomas, K. W., & Kilmann, R. H. (1974). *Thomas-Kilmann conflict mode instrument*. Tuxedo, NY: Xicom.
- Thomas, K. W., & Kilmann, R. H. (1975). The social desirability variable in organizational research. An alternative explanation for reported findings. *Academy of Management Journal*, 18, 741-752.
- Winer, B. J. (1971). *Statistical principles in experimental design*, (2nd ed.). New York: McGraw-Hill.

Received June 18, 1986 ■

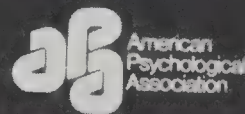
Journal of Applied Psychology

Reports investigations of interest to those doing research or working in such settings as universities, industry, government, urban affairs, police and correctional systems, transportation and defense systems, and consumer affairs. Most articles are original contributions providing new knowledge and understanding of the various fields of applied psychology (except clinical psychology); occasional theoretical and review articles are also published.

Professional Psychology: Research and Practice

Emphasizes the application and scientific support of psychology — an important subject to all psychologists. This journal publishes theoretical and data-based articles on techniques and practices used in the application of psychology.

Specifically, articles can concern research and theory on public policy as it affects the practice of psychology; current advances in applications from such fields as health psychology, community psychology, clinical neuropsychology, family psychology, and forensic psychology; standards of professional practice and delivery of services in a variety of contexts; education and training of professional psychologists at the graduate level and in continuing education; and research and theory as they concern the interests of those in the practice of psychology. *Professional Psychology: Research and Practice* also publishes brief reports.



Respected Resources in Psychology

PsycSCAN: Applied Psychology

A quarterly publication designed for the individual researcher. *PsycSCAN: Applied Psychology* contains abstracts from a group of special subscriber-selected journals. The material is derived by the PsycINFO Database from which *Psychological Abstracts* is produced. Designed for easy scanning and to help students and professionals keep up to date in their selected field.

Yes! Enter my subscription(s) to:

Journal of Applied Psychology (Quarterly)

- _____ APA Member / Affiliate Price: ~~\$30.00~~ \$24.00 (Airmail \$34.00)
- _____ Nonmember Individual Price: \$60.00
- _____ Institutional Price: \$120.00
- _____ Foreign Nonmember Individual Price: \$63.00 (Airmail \$72.00)
- _____ Foreign Institutional Price: \$127.00 (Airmail \$134.00)

Professional Psychology: Research and Practice (Bimonthly)

- _____ APA Member / Affiliate Price: \$35.00 \$28.00 (Airmail \$48.00)
- _____ Nonmember Individual Price: \$40.00
- _____ Institutional Price: \$86.00
- _____ Foreign Nonmember Individual Price: \$45.00 (Airmail \$60.00)
- _____ Foreign Institutional Price: \$97.00 (Airmail \$122.00)

PsycSCAN: Applied Psychology (Quarterly)

- _____ APA Member / Affiliate Price: \$12.50 \$10.00 (Airmail \$20.00)
- _____ Nonmember Individual Price: \$25.00
- _____ Institutional Price: \$50.00
- _____ Foreign Nonmember Individual Price: \$28.00 (Airmail \$35.00)
- _____ Foreign Institutional Price: \$57.00 (Airmail \$64.00)

TOTAL AMOUNT ENCLOSED _____

Name _____

Address _____

All subscriptions must be prepaid.
Make checks payable to the American
Psychological Association.
Mail order form to:

APA Subscription Dept.
1200 17th Street NW
Washington, DC 20036 USA

JOURNAL OF APPLIED PSYCHOLOGY MONOGRAPH

A Butterfly Catastrophe Model of Motivation in Organizations: Academic Performance

Stephen J. Guastello
Marquette University

This monograph advances and tests a model proposing that changes in performance levels, rates of absenteeism, and turnover are best described by a nonlinear interactive process that is controlled by the subject's abilities, intrinsic and extrinsic motivational factors, and organizational climate variables. In an application to academic performance, changes in grade point average from high school to college were observed for 272 freshmen at a midwestern technical university. Operationalized control variables were American College Test scores plus others selected from the Strong-Campbell Interest Inventory. Squared multiple correlation coefficients for the nonlinear (polynomial with ordinary least squares regression) hypothesis ranged from .35 to .70 for various data treatments, which were larger than values obtained for conventional linear hypotheses (.02 to .09). It is argued that (a) the theory subsumes most known motivational effects and ideas and that (b) its predictive superiority in appropriate situations warrants further motivation research of its type, plus explorations of other applications of catastrophe theory in applied psychology.

This monograph presents the development and validation of a catastrophe model of motivation in organizations. It draws on a set of concepts from differential topology, dynamic systems theory, and stochastic differential equations, as well as accepted concepts of work motivation. The arguments are developed in three steps:

1. A presentation of catastrophe theory in its general form, systems theory concepts, and statistical approaches.
2. Modeling of motivation and behavior phenomena as two-, three-, and four-dimensional catastrophe models, all of which are subspaces of the full five-dimensional (butterfly) catastrophe model of motivation in organizations.
3. The operationalization of the new theory to academic performance with empirical analysis.

Qualitatively, the resulting model incorporates many accepted motivation concepts, job satisfaction and related attitudes and process models for performance, absenteeism, and turnover. Each of those lines of research contributed plausible and distinct explanations of how an individual in an organization would be predisposed to desirable and undesirable work behavior. Several of them make a distinction between intrinsically and extrinsically motivated behavior (deCharms, 1976; Deci, 1975; Dittrich & Carrell, 1979; Dyer & Parker, 1975; Enzle & Ross, 1978; Herzberg, Mausner, & Snyderman, 1959; Mawhinney, 1979; Pritchard, Campbell, & Campbell, 1977; Wahba & House, 1974; Weiss, Dawis, England, & Lofquist, 1967), which is also germane to the purposes at hand.

New theoretical goals may pertain to formal interrelationships among hypothetical constructs and the temporal dynam-

ics of work motivation. In existing theory, individual and group differences in behavior are largely explicable by variables that simply add together; the multiplicative moderator relationship is the most sophisticated relationship typically invoked (e.g., Fishbein, 1967; Vroom, 1964). Although simplicity is preferable to complexity when other attributes of a theory are equivalent, there is reason to believe that a certain element of complexity would offer distinct advantages. For instance, work motivation and performance are stable over time under some conditions but not others. Also, the decision to quit one's job is based on a decision process over time (Landy, 1978; Lawler & Porter, 1967; Mobley, 1977; Rambo, Chomiak, & Price, 1983; Youngblood, Mobley, & Meglino, 1983). Changes in behavior may be trivial, smooth, or discontinuous; the latter would include qualitative changes. If discontinuous, the changes may be described by one of the elementary catastrophe models.

Catastrophe Theory

A Topological General Systems Theory

The central proposition of catastrophe theory is the classification theorem (Thom, 1975), which states (with qualifications) that all discontinuous changes in events can be modeled by one of seven elementary topological forms. The forms are hierarchical and vary in the complexity of the behavior spectrum they encompass. The models describe change between (or among) qualitatively distinct forms of behavior, such as remaining on a job versus quitting; they do not necessarily infer undesirable outcomes, as the word *catastrophe* connotes in common parlance. Steady states and changes in behavior are governed by one to four control parameters, depending on the complexity of the behavior spectrum under consideration. In the catastrophe theory literature the sense of control is not like that of a variable

Correspondence concerning this monograph should be sent to Stephen J. Guastello, Department of Psychology, Marquette University, Milwaukee, Wisconsin 53233.

held constant while other experimental variables are free to vary. Rather, it denotes independent variables that have a particular function in the change process. In research, the investigator would identify one or more psychological measurements that would correspond to a particular function. The term *control parameter* is additionally distinguished from the unmodified term *parameter*, more commonly recognized as the population value of a variable inferred from a sample statistic.

For theorists in many scientific disciplines, catastrophe theory offers a qualitative theory of discontinuous change and equilibria. For the applied scientist, it offers a systematic methodology for the prediction of change. The exposition of catastrophe theory is organized into four sections: the elementary cuspid models, modeling concepts, the range of its applications, and stochastic principles.

Potential Functions and Surface Equations

The elementary models are classified into two groups: the cuspoids and the umbilics. Only the former are discussed in this monograph. The elementary cuspoids involve one dependent measure, have potential functions in three to six dimensions, and response surfaces in two to five dimensions. They are the fold, cusp, swallowtail, and butterfly. The names reflect fanciful interpretations of what parts of their geometry resemble.

The response surface itself describes the range of possible behaviors, or behavioral intensities that could be observed (dependent measure). Infinitesimal or drastic changes in behavior may be observed as a function of control parameter (sets of independent measures) values. For instance ■ student may work at a steady level of proficiency in high school but receive many failing grades in college because of more than one control parameter. The general relationship between independent and dependent measures requires more than two Euclidean dimensions to describe.

Fold. The potential function for the fold is

$$f(y) = y^3/3 - ay \quad (1)$$

and its surface is defined as the set of points where

$$0 = y^2 - a, \quad (2)$$

where y is a dependent measure and a is a control parameter. The fold is the basic building block of the seven models and beyond. It describes a behavior changing from a stable state to an unstable state as a function of a . In an actual experiment, the control parameter may be operationalized as a set of independent variables.

Cusp. The cusp surface is three-dimensional and features a two-dimensional manifold (unfolding). It describes two stable states of behavior. Change between the two states is a function of two controls, asymmetry (a), and bifurcation (b). At low values of b , change is smooth, and at high values of b it is discontinuous. At low values of a , changes occur around the lower mode and are relatively small in size. At middle values of a , changes occur between modes and are relatively large. At high values of a , changes occur around the upper mode and are again small. Taken together the surface is a set of points where

$$0 = y^3 - by - a. \quad (3)$$

The cusp surface appears in the upper portion of Figure 1.

Change in the behavior of the subject is denoted by the path of a control point over time (dotted line). The point begins at some high level of a behavior, or behavior of one type, and is observed in that mode for a period of time. During that time its coordinates on a and b are changing when suddenly it reaches a fold line and drops to the lower value of the behavior, which is qualitatively different, where it remains. Reversing direction, the point is observed in the lower mode until coordinates change to a critical pair of values, at which moment the point jumps back to the upper mode. There are two thresholds for behavior change, one ascending and one descending. The shaded area of the surface is the region of inaccessibility in which very few points fall.

Bifurcation. The cusp and higher order models also have a control surface on which the bifurcation set is drawn, mapping the unfolding of the surface in (for the cusp) two dimensions. When highlighted on the response surface itself, the cusp bifurcation set induces two diverging response gradients, which are joined at a cusp point. The diverging gradients are labeled A and B on the cusp surface in Figure 1. Behavior at the cusp point is ambiguous; a classic example is the dog who sits in one position rather than attacking or running away from an unwanted stimulus (Zeeman, 1977). The diverging gradients represent varying degrees of probability that the dog might attack or run away.

In a dynamic system, the behavioral variable (y) changes values. Static systems are not change oriented; rather the observed behavior takes a few qualitatively different forms, often with gradations. Static bifurcations underlie discontinuous or multimodal probability density functions; the bimodal distribution of crude birth rates for various nations is such an example (Cobb, 1978).

Swallowtail. The swallowtail describes a behavior spectrum consisting of two stable modes plus two unstable areas. Its surface defined as the set of points where

$$0 = y^4 - cy^2 - by - a. \quad (4)$$

The new control parameter c is called a *bias* or second asymmetry factor.

Butterfly. The butterfly surface is trimodal, describing a spectrum of three qualitatively different behavioral outcomes. Once again there are regions of inaccessibility among the modes. The function is five dimensional, and what appears in Figure 1 is the most interesting three-dimensional sectioning. The surface features a pocket, which allows the control point (denoting which behavior is taking place) to slip through to the middle mode, or oscillate between the top and bottom modes. It is defined as the set of points where

$$0 = y^5 - dy^3 - cy^2 - by - a. \quad (5)$$

The new control parameter d is called the *butterfly* or second bifurcation factor.

Other sectioning schemes can be visualized as well. The manifold and bifurcation set for the butterfly is four dimensional, and a two-dimensional sectioning is shown in the upper right portion of Figure 1. For a full series of sectionings see Poston and Stewart (1978a), Saunders (1980), Thompson (1982), Woodcock and Davis (1978), and Zeeman (1977). There are four diverging gradients, two of which are positive and two are negative.

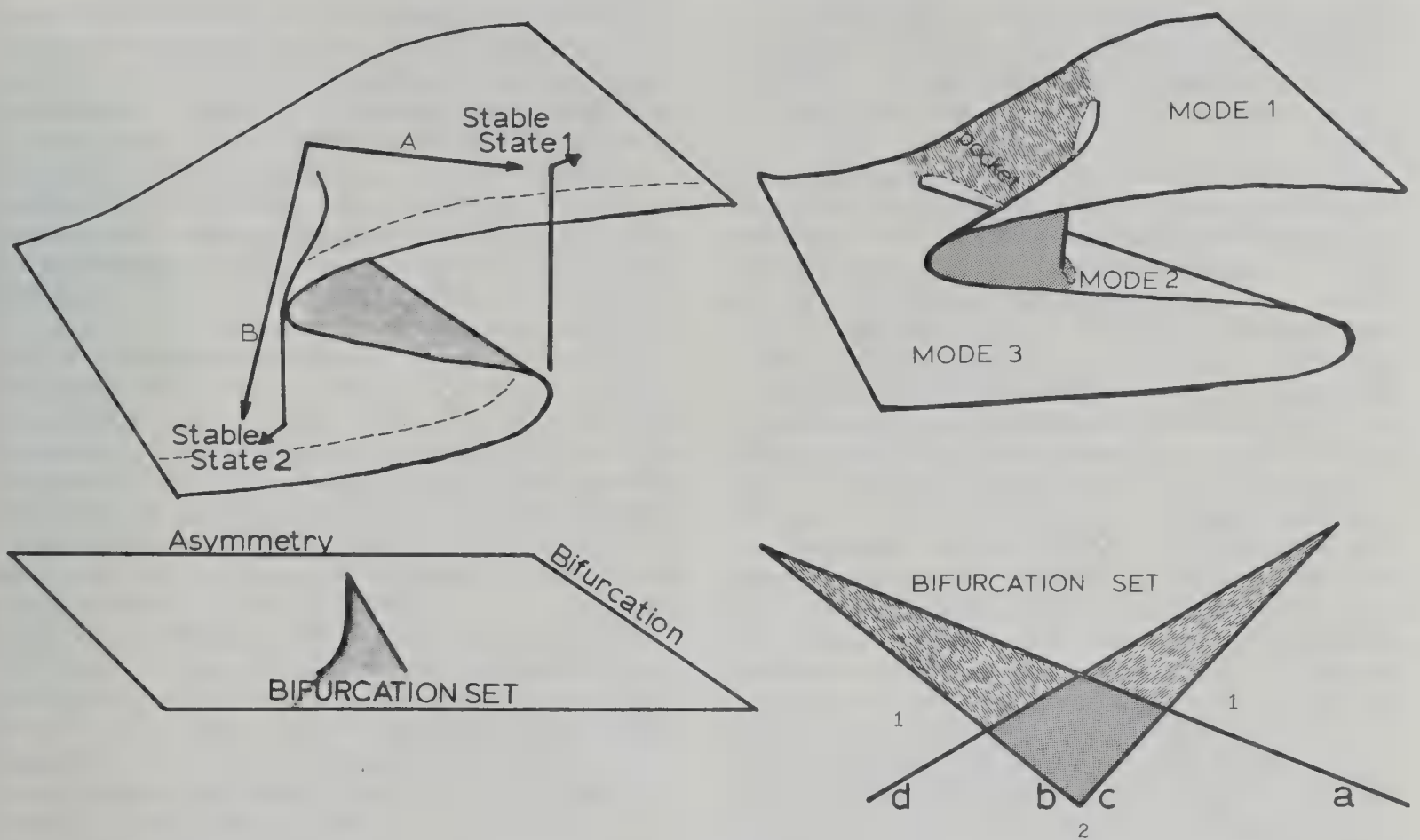


Figure 1. The cusp (left) and butterfly (right) catastrophes with their respective bifurcation sets. Shaded portion of the cusp surface is the zone in which very few points fall; its shadow on the control plane is the bimodal behavior zone, where extreme high and low scores are observed. The heavily shaded area on the butterfly bifurcation plane is the zone of trimodality. Grained portions are zones of bimodality between adjacent modes. White area marked "2" is a zone of bimodality between extreme modes. White areas marked "1" are unimodal.

The set of points comprising the cusp surface exist in Cartesian space defined by axes y , a , and b . As more variables are added to form a butterfly, however, the five axes become organized at 60° to each other, rather than at 90° (Lu, 1976).

Classification. The elementary cuspoids (as well as other nonelementary models) vary in complexity. Complexity is signified by the highest exponent for the behavioral variable in the surface equation, for example, the quadratic term for the fold, the cubic for the cusp, and so forth. The leading exponent denotes the number of control parameters, each with its own unique function, and the complexity of the behavioral array they perpetrate.

Modeling Concepts

There are seven key steps in the process of model building. (a) Identify whether the system is a static or dynamic process. This distinction was discussed above. (b) Identify the nature of exposure, whether the exposure is a random or systematic random force (see below). (c) Identify the qualitative behavioral modalities, stability patterns, and any gradations that might be observed. (d) Determine which of the seven models, if any, is appropriate, and hence how many control parameters are required; it is sometimes productive to begin with the cusp before

proceeding to a more complex model. (e) Identify by observation or pre-existing research reports exogenous variables that behave in the system as one of the control parameters. (f) Determine, if possible, what combinations of these variables are responsible for gradient behavioral contours. (g) Make predictions from the model and choose an appropriate experimental design to test it.

Random force. A random force occurs when a system is exposed to an entropy-inducing set of events, in which each element of the system (e.g., subjects in the experiment) has an equal probability of exposure (Agu, 1983). As elements increase in entropy (or randomness), the probable location of each in space expands to the limits of its confinement. Bifurcation serves to reduce entropy by partitioning energized elements into neighborhoods of relative stability and equilibrium (Thompson, 1982).

One can next define situations where elements are not equally exposed to random force, but instead are systematically exposed. The level of exposure (or experimental variables governing level of exposure) becomes the b parameter in the surface equation for the cusp. Where there is more than one parameter of exposure, c is introduced, and the model becomes a swallowtail. In a swallowtail, b and c can coact only additively or only subtractively. If both can occur, then butterfly factor d governs

the degree to which b and c are additive or subtractive. In all cases, control a governs the proximity of the case to the critical manifold. When random force and bifurcation are considered jointly, the coaction of the two may be said to cause a phenomenon.

Equilibria and stability. It should be recognized that the catastrophe models comprise a highly evolved subset of bifurcation and equilibrium phenomena (Fararo, 1978; Thompson, 1982). Some general concepts apply here and in linear and nonlinear dynamic systems theory more generally. The path of a control point over some region of a response surface is called a *trajectory*. If an equilibrium exists, and it is stable, a control point originating near the equilibrium position will remain near it. The control point may approach it asymptotically, or else orbit around it. If an equilibrium is unstable, trajectories approaching it will not remain in its neighborhood.

Stable equilibria, like the two modes of a cusp catastrophe, exert an attractor force on the control point; an attractor force may be viewed as a gravitational force. The space between equilibria, corresponding to the zone of inaccessibility or a statistical antimode, exerts a repeller force on the control point. The cusp point and corresponding points in other models, collectively known as points of degenerate singularity, are unstable.

A globally stable situation is one in which asymptotic stability occurs no matter how far from the equilibrium the control point is allowed to deflect. *Local stability* names the situation in which the control point may jump from orbit around one equilibrium to another and not return (unless values of control parameters change) and thus orbit the latter equilibrium. Changes in behavior along a catastrophe model surface are reversible. *Hysteresis* is the phenomenon by which behavior oscillates between stable states. The term also denotes the double threshold effect that is observed in such cases.

Range of Application

The elementary catastrophe models have been used for problems from a wide range of physical and social science disciplines. Pure mathematical treatments were considered first (Arnol'd, 1974, 1981; Callahan, 1980, 1982a; Chilingworth, 1976; Gibson, 1979; Gibson, Wirthmuller, du Plessis, & Looijenga, 1976; Lu, 1976; Poston & Stewart, 1976, 1978a; Stewart, 1980, 1981a; Thom, 1975; Zeeman, 1977). Physical science applications include thermodynamic phase transitions, structural mechanics and the buckling of elastic beams, the stability of ships and oil rigs, the stability of aircraft, optics, climate, and geophysical catastrophes (Dold & Eckman, 1976; Gilmore, 1981; Poston & Stewart, 1978a; Saunders, 1980; Sinha, 1981; Stewart, 1981b, 1982; Thom, 1975; Thompson, 1982; Woodcock & Davis, 1978; Zeeman, 1977).

Biological applications address denaturation phenomena (Benham & Kozak, 1976, 1978), embryological development (Thom, 1975; Woodcock, 1978), immunological systems (Merrill, 1980a, 1980b), thyroid dysfunction (Cobb & Zacks, 1983), and brain functioning (Thompson, 1982; Zeeman, 1977). In a recent trilogy (Tanyi, 1982a, 1982b, 1982c) the set of seven models described a series of biochemical reactions, the relationship of that system to macromolecular self-organization, a food chain, and the Lotka-Volterra functions for prey-predator relationships. The Lotka-Volterra models for two- and three-species

competition and prey-predator relationships have in turn provided theoretical substrates for game theory and urban planning catastrophes (Dendrinis, 1980; Dendrinis & Mullaly, 1983; Haag & Dendrinis, 1983; Kapur, 1981; Poston & Stewart, 1978a; Schofield, 1980; Thompson, 1982; Wilson, 1981).

Business and economic catastrophes have spanned the stability of stock prices (Balasko, 1978; Weintraub, 1983; Zeeman, 1974), oil prices (German, 1983/1984), market planning (Chidley, Lewis, & Walker, 1978; W. Smith, 1984), dispersions of insurable risks (Guastello, 1984a), business activity cycles (W. Smith, 1980), and operations research (Carhart, 1984).

Besides the bioecological models mentioned above, catastrophe models have been found to describe the recurrence of spruce budworm outbreaks (Casti, 1982), prison riots (Guastello, 1982a; J. Smith, 1980; Zeeman et al., 1976), crime rates (Jiobu & Lundgren, 1978), and other disturbances of internal social order (DeGreene, 1978; Schofield, 1977). Of sociological importance are models for the emergence of urban slums (Dendrinis, 1979) and the distribution of crude birth rates across nations (Cobb, 1978). At the political level, juntas in Latin America during the past 20 years (Adelman & Hihn, 1982), the utility of waging war (Morrow, 1983; Zeeman, 1977) and arms race dynamics (Bosserman, 1982; Fararo, 1978) all follow catastrophe rules of advance.

Applications in psychological theory include psychophysical judgments (Ayers, 1981; Yelen, 1980) with application to clinical decision making (Lueger, 1982), parameterization of learning curves (Baker & Frey, 1980; Frey & Sears, 1978), perception of optical illusions (Poston & Stewart, 1978b; Stewart & Pergey, 1983), attitude change and behavior (Cobb, 1980; Flay, 1978), reactance to social feedback (Tesser, 1980), the treatment of anorexia nervosa (Callahan, 1982b; Zeeman, 1977), color vision under stressful conditions (Guastello, 1982b), progression through stages of cognitive development (Saari, 1977), and sexual preference (Evan & Zeiss, 1984). Of a more applied nature, models have been proposed for the effect of alcohol on driving speed (Cobb, 1981a; Zeeman, 1976) and other traffic calamities (Dendrinis, 1980; Furutani, 1976a, 1976b, 1977), status displays for the stability of power generators (Sallam & Dineley, 1983), labor-management negotiations (Oliva & Capdevielle, 1977; Oliva, Peters, & Murthy, 1981), equity in organizations, work performance, and absenteeism (Guastello, 1981, 1984b, 1984c), turnover (Abelson, 1982; Guastello, 1981; Sheridan, 1980, 1985; Sheridan & Abelson, 1983), induction and decisions (Dockens, 1979; Johnson, 1982; Keown, 1980), the reliability of confidential information (Fung, 1980), organizational reactance to change (Bigelow, 1982), two-stage personnel selection and training evaluation (Guastello, 1982c), physical fatigue (Guastello & McGee, 1982), performance decrement under increased load (Guastello, 1985a, 1985b), and occupational safety and accidents (Guastello, 1984d; Guastello & Dizadji, 1984).

Many physical science applications were identified through mathematical proof. In such cases there still remained the question of whether translating a phenomenon from one system to another accomplished anything useful. In thermodynamics, the cusp catastrophe made predictions equal to those of the Van der Waals equations for phase changes. The added value was that cusp geometry is a subset of butterfly geometry; the butterfly model explained and made predictions for complex change

phenomena that were not covered by the original theory (Gilmore, 1981). Similarly, cusps for beam buckling and ship stability could be expanded to a wigwam (nonelementary cuspid) and applied to an unsolved problem in small aircraft stability.

Social science applications are typically substantiated on the basis of their ability to solve a conceptual problem not answered by existing theory, synthesize several conceptual elements, or improve experimental accuracy (Bruter, 1978; Guastello, 1981; Oliva & Capdevielle, 1980; Poston & Stewart, 1978b; Stewart & Peregoy, 1983; Sussmann & Zahler, 1978a, 1978b). The testing of social science theory relies heavily on statistical inference. The concept of statistical evaluation for topological hypotheses had been pronounced a contradiction in terms for years (Stevens, 1951; Woodcock & Davis, 1978). Several statistical models were eventually proposed, based on double-normal (Benham & Kozak, 1976), double-gamma (Gilmore, 1981), binomial and chi-square (Jiobu & Lundgren, 1978; Sheridan, 1980), double binomial (Oliva, Peters, & Murthy, 1981) distributions, Bayesian techniques (J. Smith, 1980; J. Smith, Harrison, & Zeeman, 1981), and the more general theory of stochastic differential equations (Cobb, 1978, 1980, 1981a, 1981b, 1981c; Cobb, Koppstein, & Chen, 1983; Cobb & Watson, 1980). Statistical catastrophe theory, which is the latter approach, is the strongest of the group because it is based on probability density functions specific to the cusp, butterfly, or other models. The polynomial regression equations that proceed from it offer considerable experimental design flexibility (Guastello, 1982b, 1982c).

Statistical Catastrophe Theory

Probability density functions. The first proposition is that any function of y can be transformed into a probability density function (pdf) of y by the formulation

$$\text{pdf}(y) = \xi \exp \left[-\int f(y - \lambda)/\sigma \right] \quad (6)$$

In Equation 6, λ is the lower limit of behavior variable y , σ is a measure of variability (which need not be an ordinary standard deviation, see below), and ξ is a constant introduced to ensure unit density. Thus, the cusp pdf is represented by the function

$$\text{pdf}_c(y) = \xi \exp \left[-z^4/4 + bz^2/2 + az \right], \quad (7)$$

where

$$z = (y - \lambda)/\sigma \quad (8)$$

(Cobb, 1978, 1981a, 1981b, 1981c). The parameter σ is dispersion around the modes rather than dispersion around the mean as in the ordinary standard deviation.

The cusp pdf is three-dimensional. At low values of b it is unimodal, resembling a normal distribution. At higher values, the pdf becomes increasingly bimodal. It is not difficult, therefore, to define the butterfly pdf as follows:

$$\text{pdf}_b(y) = \xi \exp \left[-z^6/6 + dz^4/4 + cz^3/3 + bz^2/2 + az \right]. \quad (9)$$

The resulting pdf is five-dimensional with a unimodal region, two bimodal regions, and a trimodal region.

Parameter estimation. To evaluate whether a given distribution of behavior scores are truly cusp catastrophic in nature,

one uses an iterative parameter estimation procedure (Cobb, 1980, 1981a, 1981c). There are $2k + 4$ parameters to be estimated (8 for the cusp), which result in values for the six variables in the following stochastic cusp equation:

$$0 = (a_1 + a_0) + (b_1 + b_0)z - z^3. \quad (10)$$

In Equation 10, z is defined as before, a_1 and a_0 denote the range of the latent asymmetry parameter, and b_1 and b_0 the range of the latent bifurcation parameter. Having estimated the parameters of the latent pdf, one can then compute an r^2 coefficient for goodness of fit.

The use of r^2 for goodness of fit has been dubbed "pseudo- r^2 " (Cobb, 1981b) because it does not permit hypothesis testing in the usual sense. To test hypotheses concerning experimental variables in control parameter positions one must proceed a step further by estimating latent a and b parameters and correlating them with exogenous experimental variables. If these correlations are statistically significant, then it is concluded that the experimental variable contributed to the latent parameter. Pseudo- r^2 by itself evaluates the cusp structure only.

As a final step, R^2 is calculated for the explicit regression equation

$$y = \beta_0 + \beta_1 a + \beta_2 b \quad (11)$$

to compare the linear-only hypothesis with pseudo- r^2 for the cusp hypothesis. Equation 11 is explicit because it takes the form $y = f(a, b)$ rather than $0 = f(y, a, b)$, which is implicit.

Parameter estimates may be calculated by the methods of moments, by approximation theory, and by maximum likelihood; proof is also available that least-squared solutions are produced (Cobb, 1981c). A FORTRAN program is available for cusp parameter estimation (Cobb, 1980). Parameter estimation procedures extend in theory to fold, swallowtail, and butterfly models.

Dynamic difference equations. The parameter estimation theory is most pertinent to the evaluation of static catastrophe distributions. The dynamic difference equations proceed from the same theory, and apply where a behavior is measured at two points in time. They were first suggested by Cobb (1978, 1981b) and later applied to problems of industrial psychological relevance (Guastello, 1982b, 1982c).

Any of the deterministic catastrophe surface equations can be converted to a statistical expression by replacing 0 with Δ_z , and inserting empirical weights among the terms. Hence the statistical cusp-difference model is

$$\Delta_z = \beta_0 + \beta_1 z_1^3 + \beta_2 z_1^2 + \beta_3 b z_1 + \beta_4 a. \quad (12)$$

In Equation 12, λ and σ are pre-estimated separately at each point in time. The sample standard deviation (ordinary) is used as the measure of dispersion. The lower limit is the lowest value of y for the sample or an assumed 0 if $\lambda \geq 0$. The objective is to ensure true $z \geq 0$. Controls a and b are similarly normalized. R^2 coefficients of the ordinary least squares variety are produced, which can be calculated on standard statistical packages.

A network of hypotheses is contained in Equation 12. First, there is the overall statistical significance of the model. The R^2 indicates the degree to which the data fit the cusp geometry. F (or t) tests on beta weights denote the contribution of each term in the model. In optimal situations each term in the model accounts for a unique portion of criterion variance. The power

potential denotes structure and implies a particular bifurcation system.

Under less than optimal conditions bivariate correlations between the individual terms and the difference criterion may be significant, but empirical weights may not all be significant in the multiple model. Here the validity of the model is still upheld and a cross-validation strategy could confirm the conclusion. If doubt still remains, a fold model should be constructed, tested, and compared to the cusp results.

The quadratic term in Equation 12 requires some explanation. The mathematical cusp model is written without the quadratic term because its significance is dependent on a change in λ ; the catastrophe models are robust with respect to changes in λ and σ . Changes in σ depend on β_1 , and changes in λ depend on β_2 . If, after normalizing y , there still remains a nonzero constant λ , the introduction of z_1^2 restores the empirical model to canonical form. This is the Tschirnhaus transformation (Poston & Stewart, 1978a, p. 442); the introduction of the new term adjusts the weights of the remaining terms. If, on the other hand, λ becomes a factor, then the next higher dimensional catastrophe is implied (Cobb, 1981b). Similarly, the butterfly difference equation contains a quartic term,

$$\Delta z = \beta_0 + \beta_1 z_1^5 + \beta_2 z_1^4 + \beta_3 dz_1^3 + \beta_4 cz_1^2 + \beta_5 bz_1 + \beta_6 a. \quad (13)$$

It should be made clear that the catastrophe difference equations are a type of time series analysis that is neither widely known nor widely accepted. There are no published theorems on the asymptotic properties of the estimators, nor are there any published simulation studies on the subject. There are no known discordancies with standard regression theory, however.

Utility coding the criterion. There are two types of variance in a catastrophic distribution: between-mode variance and within-mode variance. When within-mode variance decreases to none, one obtains a situation conventionally addressed by discriminant analysis (DA): Define variables that maximally separate qualitatively defined groups. In DA there are no theoretical provisions for change dynamics and equilibria, and it must therefore provide a weaker descriptive solution when the phenomenon is truly a catastrophe.

According to Novick (1980) there are statistical situations in which certain ranges of scores on a variable of interest have a qualitatively different meaning than in others. In such cases it is recommended that the behavior index be multiplied by a 0–1 or other utility variable. The use of utility vectors has also been incorporated into some catastrophe design modifications.

The bimodal criterion in the cusp catastrophe is continuously valued. In cases where there are qualitatively different score regions, rather than a statistically obvious bimodal density, the fit of the cusp model can be dramatically enhanced by utility coding the criterion. Utility coding schemes are of course equally applicable to butterfly and other models. For a distribution of difference scores, positive versus negative differences can be construed or defined as qualitatively different modes.

Three types of coding schemes have been used in conjunction with catastrophe problems: the utility (or bimodal) transformation, and the upper and lower implicit discriminant functions, (IDF; Guastello, 1982c). For the bimodal transformation, scores below -1 standard deviation are recoded to 0.00, and others are simply multiplied by 1.00. For the upper IDF, scores below $+1\sigma$ are recoded to 0.00.

The lower IDF, on the other hand, calls for scores above -1σ to be recoded to 0.00. This situation is one of selecting cases out of the system rather than into the system. The lower IDF is a retrograde transformation: Resulting positive difference scores denote negative change, and resulting negative difference scores denote positive change. In addition large differences denote close proximity to the threshold, whereas small differences denote longer movement away from the threshold. Thus, the lower IDF accentuates cases in the critical zone around the change line.

IDFs and utility codes dramatically enhance prediction or give no information of additional use. Imagine that the data points are suspended in three- (or five-) dimensional space. They can be perfectly or imperfectly aligned with the latent surface. If perfectly aligned, the R^2 will be largest for the uncoded criterion. If, on the other hand, the unfolding of the latent surface occurs in the upper range of scores or the lower range, one of the IDFs or utility coded schemes will produce better results. "Better" is defined in terms of overall R^2 and interpretability of terms contributing to control parameters.

Experimental comparison. R^2 for the butterfly hypothesis is compared to R^2 for the following linear-only hypothesis:

$$\Delta y = \beta_0 + \beta_1 a + \beta_2 b + \beta_3 c + \beta_4 d. \quad (14)$$

In many cases the following alternative pre-post linear hypothesis is a viable comparison:

$$y_2 = \beta_0 + \beta_1 y_1 + \beta_2 a + \beta_3 b + \beta_4 c + \beta_5 d. \quad (15)$$

Equation 15 differs from 14 only in that y_1 is moved from the left to the right side of the equation and given an empirical weight of its own. In a few studies on record where the catastrophe hypothesis was tested using the difference equations or Cobb's procedure, the R^2 for the catastrophe model is dramatically larger than R^2 for the linear model, such that additional significance tests are superfluous (Cobb, 1981a; Guastello, 1982a, 1982b, 1982c, 1984c).

Psychometric properties of difference scores. Difference measurements in psychology are traditionally less preferred than additive or simple ones. In effect, when the difference between two measures is taken, true scores are subtracted but errors are added (Lord & Novick, 1968). The essence of the proof rests on the proposition that errors are uncorrelated with true scores or other errors. There is other evidence, however, that errors in difference scores representing intraindividual change are correlated from one time frame to the next (Labouvie, 1980). Such dependent errors are thought to be accounted for in catastrophe models by the power potentials (Cobb & Zacks, 1983; Guastello, 1982a). Some random error in measurement would still exist.

The dynamic difference equations provide the strongest inferences of the various statistical models currently available. Strength of inference is measured in terms of the similarity of a data distribution's properties to the deterministic cusp shape, the number of parameters one must estimate, and quality of the prediction equation obtained. As one might expect, if the data requirements cannot be met, one may opt for another model. For instance, a butterfly difference equation involving only one experimental independent variable per control parameter would require seven estimated regression parameters and

would deliver a prediction model in the form of one equation. The implicit parameter estimation procedure (Cobb, 1981a), on the other hand, would require 12 estimated parameters for the basic equation, plus others to associate exogenous variables to latent control parameter estimates; prediction of the dependent measure would be cumbersome at best. The parameter estimation procedure is amenable, however, to testing *static* hypotheses where the dependent measure is a single observation rather than a difference score. The bimodal alternatives can accommodate a discrete qualitative dependent measure but are based on a weaker notion of the underlying pdf and provide no overall prediction equation. The remaining alternatives often involve a relatively long time series for a dependent measure and have been used only in situations least resembling psychological problems.

Markov assumption. The catastrophe surface models assume that a perfect Markov process is operating, meaning that the control point has no memory of where it came from beyond the prior instant. Depending on the specifics, violation of the assumption could result in flattening the surface (dampening catastrophic fluctuations) or expanding the dimensionality of the surface by introducing additional convolutions.

Butterfly Model of Motivation on Organizations

Cusp, fold, or swallowtail subspace models are described below, and integrated into the butterfly model of motivation in organizations as follows: two-stage personnel selection, the Yerkes-Dodson rule, stability of performance, opponent process theory, models for turnover and absenteeism, and approach-avoidance gradients.

Motivation-Performance Subspaces

Two-stage personnel selection. The implicit theory in use for selection decisions is that there is a linear relationship between ability (or some comparable measure) and performance on the job. Motivation is often thought to intervene as a moderator variable (Edwards & Waters, 1981; Locke, Mento, & Katcher, 1978); this idea is similar to the expectancy theory proposition, that is, that performance is a multiplicative function of motivational force and ability (Vroom, 1964).

In many situations, job candidates are hired and subjected to a training program or trial period before being permanently retained with the organization. Theoretical models for two-stage decisions are sparse (Cronbach & Gleser, 1965). The cusp model for the situation (Guastello, 1982c) bears little resemblance to earlier proposals. The exposure to the training program or trial period is thought to induce a bifurcation effect by which candidates reach acceptable performance levels (upper mode) or leave the organization by choice or management decision (lower mode). Where all candidates are given uniform exposures, demographic group membership may be responsible for a bifurcation effect instead. Similarly, individual differences in motivation or task interest may contribute to parameter b .

Ability is the asymmetry variable. Note that when $z_1 = \lambda$ for all cases, the cusp-difference equation reduces to $z_2 = \beta_0 + \beta_1 a$, which is the simple predictive validity model for the ability measure. The cusp model for two-stage selection was validated in simulation form, for which catastrophe R^2 coefficients

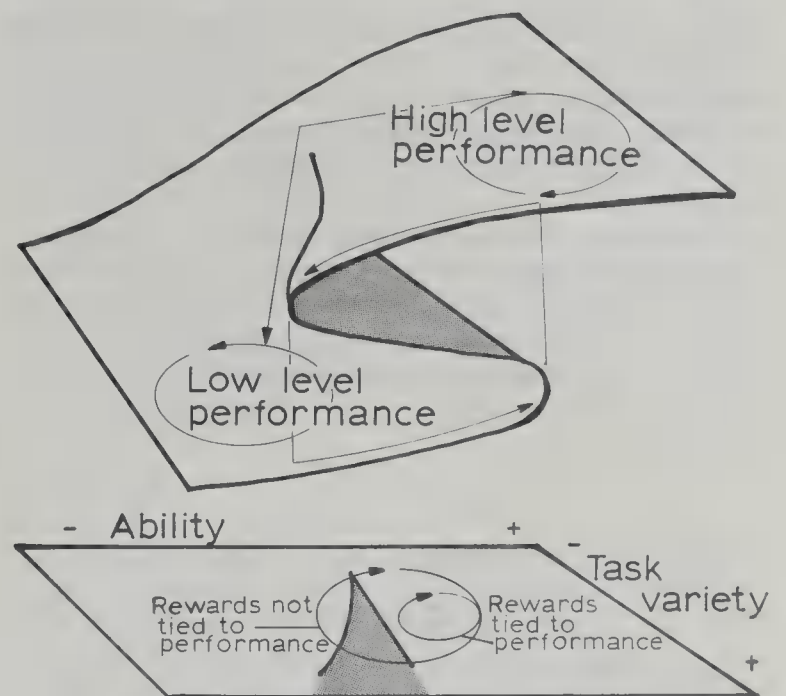


Figure 2. Cusp model for two stable performance levels with paths of the control point drawn on the response surface. (For cases of high task variety but only moderate ability, behavior changes discontinuously between the two stable states. Other orbits of interest are shown on the bifurcation plane. A large variability in performance is observed where rewards are not tied to performance.)

ranged from .36 to .95, compared to .04 for concurrent validity (Guastello, 1982c).

Yerkes-Dodson Rule. The inverted-U relationship between motivation, arousal or anxiety, and performance can be decomposed into two linear functions (McGrath, 1976): First, higher levels of motivation lead to higher levels of performance. Second, performance decrement follows increased task difficulty. Task difficulty is logically a negative function of ability. These effects are modeled by trajectories from the cusp point to the locally stable orbits on the cusp surface (see Figure 2).

Stability of performance. Work performance in industry is more stable over time under some conditions than under others (Rambo et al., 1983). For instance, performance is more variable when rewards are not tied to performance than when they are tied to performance. Also, performance is more variable when the job requires a greater task variety compared to single task conditions. These effects are modeled by orbits drawn on the cusp bifurcation set, which indicate the path of the control point on the surface (see Figure 2). The former situation applies to extrinsic motivation (hourly pay versus piecework). The latter situation concerns intrinsic motivation, specifically interest value of the task, and the theory of job design advanced by the Herzberg et al. (1959).

For the situation of pay tied to performance, let pay be somewhat motivating, that is, $b > 0$ in the cusp, and the ability of subjects be adequate. Next, let motivation decompose into bifurcation gradients—positive where performance is obviously rewarded, and negative where it is not. When pay is tied to performance, performance varies only a small amount over time, orbiting the upper mode. When it is not tied to performance, variability between modes occurs.

Task variety can be positively or negatively motivating. In the

positive case, task variety enhances motivation by providing the employee with a range of responsibilities, which add up to a whole job rather than just a fragment of the operation. On the other hand, if the employee is given a group of “dumb jobs” the impact would be demotivating since “they just want more work out of me.”

Rambo et al. (1983) showed that performance stability approaches an asymptotic limit. A formula was adopted from Kessler and Greenberg (1981) for “structured change.” The covariance of performance between two time periods is equal to the sum of two components; performance at Time 1 squared, and the covariance of performance at Time 1 with change in performance. That is,

$$S_{x_1x_2} = S_{x_1}^2 + S_{x_1\Delta x} \tag{16}$$

Equation 16 states that the following is an optimal relationship:

$$\Delta y = \beta_1 y_1^2 + \beta_2 y_1 \tag{17}$$

Because job performance is typically a linear function (albeit an imperfect one) of job-related ability, one can substitute ability, *a*, for *y*₁, and the result is the statistical difference equation for the fold catastrophe (Guastello, 1982b, 1982c), which is

$$\Delta z = \beta_0 + \beta_1 z^2 + \beta_2 a \tag{18}$$

The only remaining discrepancy between Equations 17 and 18 is the correction for location and scale, which is inconsequential to the relationship of interest.

Rambo et al. (1983) found that the correlation between performance measures from two points in time reached an asymptotic lower limit as the number of weeks between measurements increased from 1 to 178. During that time, working and incentive conditions did not change. Computed values of structured change were found to be fairly constant throughout the time series. The dynamic systems theory interpretation of the results, therefore, is that the work situation was globally stable, deviations are fold catastrophic, and the orbital period is about 30 weeks. The specific results did not specify the equilibrium value of performance, but that some equilibrium was approached. Not much change occurred between consecutive performance weeks; hence, high correlations were observed. The elbow of the correlation-lag time curve broke at about 20 weeks, beyond which time all variation that was likely to occur

did occur. Because the asymptote was calculated for all possible pairs given lag length, global stability is inferred. Global stability is additionally reinforced by knowledge that the work situation was not manipulated during the time horizon.

Opponent processes. The opponent process theory of job satisfaction describes an oscillation between two qualitatively different levels of performance (Landy, 1978), thus indicating a fit for a cusp catastrophe model. The bifurcation gradients may be the Herzberg et al. (1959) motivator and hygiene factors, or intrinsic and extrinsic factors, which are in turn thought to be linked to sympathetic and parasympathetic nervous system processes. The greater the number of hygiene and motivator elements, the larger the possible variation in performance. The second part of the process calls for a decline in peak performance over time, which is caused by boredom. Where ability is held constant, peak decline contributes to the asymmetry parameter. Six steps in the opponent process phenomenon are modeled as trajectories on a cusp surface in Figure 3.

When ability is free to vary, moderately high-ability persons would show the peak decline, or boredom effect. Low-ability subjects would simply show poor performance with only small variability. Persons of high ability, though they may experience vast amounts of boredom, can more easily keep their performance levels from dropping too low. Boredom can be considered as negative intrinsic motivation. It has the effect of shifting the peaks to a new mode as an asymmetry parameter would.

The effects of ability, hygiene factors (extrinsic motivation), and boredom (intrinsic motivation) can be next combined into a swallowtail model. The surface has two stable modes for acceptable but uninspired performance, and one for superior feats. The unstable zones represent organizationally unacceptable performance. The control are as follows: *z* = ability, *b* = hygiene factors, and *c* = cycle number or boredom (see Figure 4).

The six steps in the opponent process appear on the swallowtail surface. The surface is four-dimensional and requires sectioning for cases where *a* = 0.0 and cases where *a* > 0.0. At Step 4, the control point drops through the inaccessible region to the unstable mode and returns to the lower stable mode at Step 5. The swallowtail bifurcation set is also shown for comparison with other models. Additional implications of the catastrophes

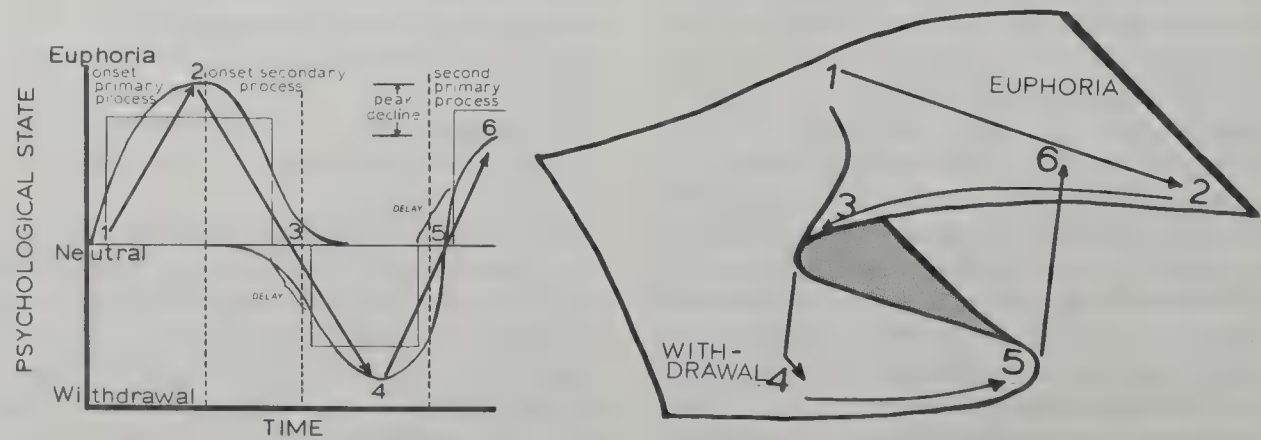


Figure 3. The generalized opponent process model with cusp catastrophe reinterpretation. (Key steps in the process are represented by corresponding trajectories along the cusp surface.)

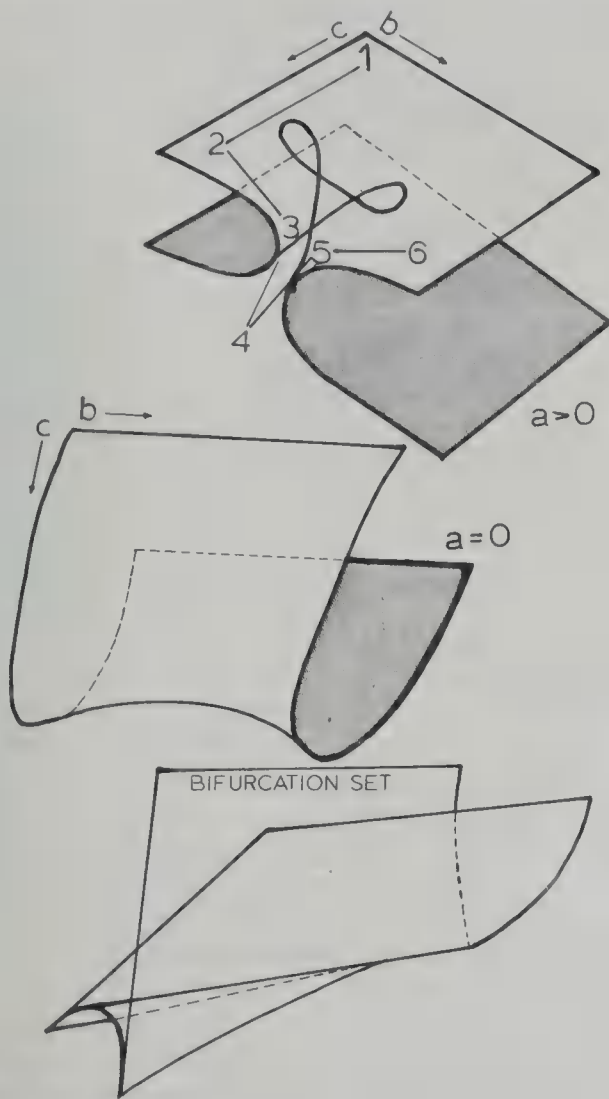


Figure 4. The six steps of the opponent process are shown as trajectories on a swallowtail surface. (Peak decline occurs at Steps 5 and 6.)

for opponent process models are treated separately elsewhere (Guastello, 1984b).

Turnover and absenteeism. Motivation and satisfaction theories are as relevant to turnover and absenteeism phenomena as they are to work performance. It is generally true that absenteeism is precipitated by a lack of intrinsic satisfaction, and turnover is precipitated by a lack of both intrinsic and extrinsic elements (Dittrich & Carrell, 1979; Locke, 1976). Because these forms of motivation and satisfaction are already implicated in the model building process, a catastrophe model for inspection must also address absenteeism and turnover to some extent, particularly in light of the discontinuities in work behavior that are involved. Indeed, such catastrophe models have been proposed and have withstood preliminary tests (Guastello, 1981, 1984c; Sheridan, 1980, 1985; Sheridan & Abelson, 1983). Abelson (1981/1982) expanded the cusp concept by identifying stages of the Mobley (1977) process as positions along the cusp surface.

Approach-avoidance. If one were to view the cusp surface divergence gradients in two dimensions, they would be of equal length but different slopes; the gradient labeled in Figure 1 is steeper. This relationship describes the differential approach and avoidance gradients that have long been observed in animal

and human social behavior. The subject avoids a negative stimulus faster than it approaches a positive one (Brown, 1948; Wegner & Vallacher, 1977). The approach-avoidance gradients are implicit in all motivation-performance subspaces, and account for the differential impact of positive and negative utilities in expectancy (Leon, 1981). Subjects were on an upper performance mode at the time their utilities were measured. Slight negative influences would not result in a negative performance deflection. Only when the negative influences are sufficiently large (relative to the subject's initial proximity to the upper equilibrium) is a negative performance deflection observed.

Butterfly Model

It is now possible to state a general hypothesized butterfly model for motivation and performance dynamics in organizations. The performance subspace models just described are all incorporated, along with implications for absenteeism and turnover. Any particular application of the model, including the one that ensues, would demonstrate several, but not necessarily all of the hypothesized features.

Surface. Two stable, qualitatively different modalities of work performance were described above in conjunction with the opponent process model. The turnover cusps (Sheridan, 1980, 1985; Sheridan & Abelson, 1983) simply contrasted staying on the job with quitting. Merging the two ideas, there are three distinct modes of performance: high, good enough to get by, and poor enough to warrant termination. A butterfly surface would be required to describe change in behavior, or the distribution of behavior, among the three modes. The butterfly also allows for voluntary turnover from good performers. In studying change in behavior among the two work modalities and turnover, turnover would be scored 0.00 on a performance scale ranging from 0.00 to some positive value (Guastello, 1981, 1982c, 1984b).

Absenteeism can be thought of as a hysteresis between staying on the job and quitting. Frequent absences can become a stable mode of behavior in its own right (Guastello, 1981). Although it is not often mentioned in the literature, three modes of absence can be typically identified: those persons who gravitate to no absences at all, those who gravitate toward a perceived average level, and the chronic absentees. The modalities for absenteeism, performance, and turnover can be organized into a butterfly response surface as follows.

On the upper mode, subjects would show self-directed, internally committed behavior: high output and high quality work. Innovation, which would be partially based on prerequisite abilities, would occur at the extreme end of this subdivision. Absenteeism rates would gravitate toward virtually none, although, conceivably, an internally directed and competent person might organize their work to permit an occasional day off. Some subjects would harbor a strong intent to leave the organization, while others would harbor none. No discernible difference in the work behavior between these two groups would be expected.

The middle equilibrium is thought to be characterized by externally motivated behavior at low levels of commitment. Innovation would not occur for high ability subjects. Quantity and quality would be merely adequate. Absenteeism would occur in

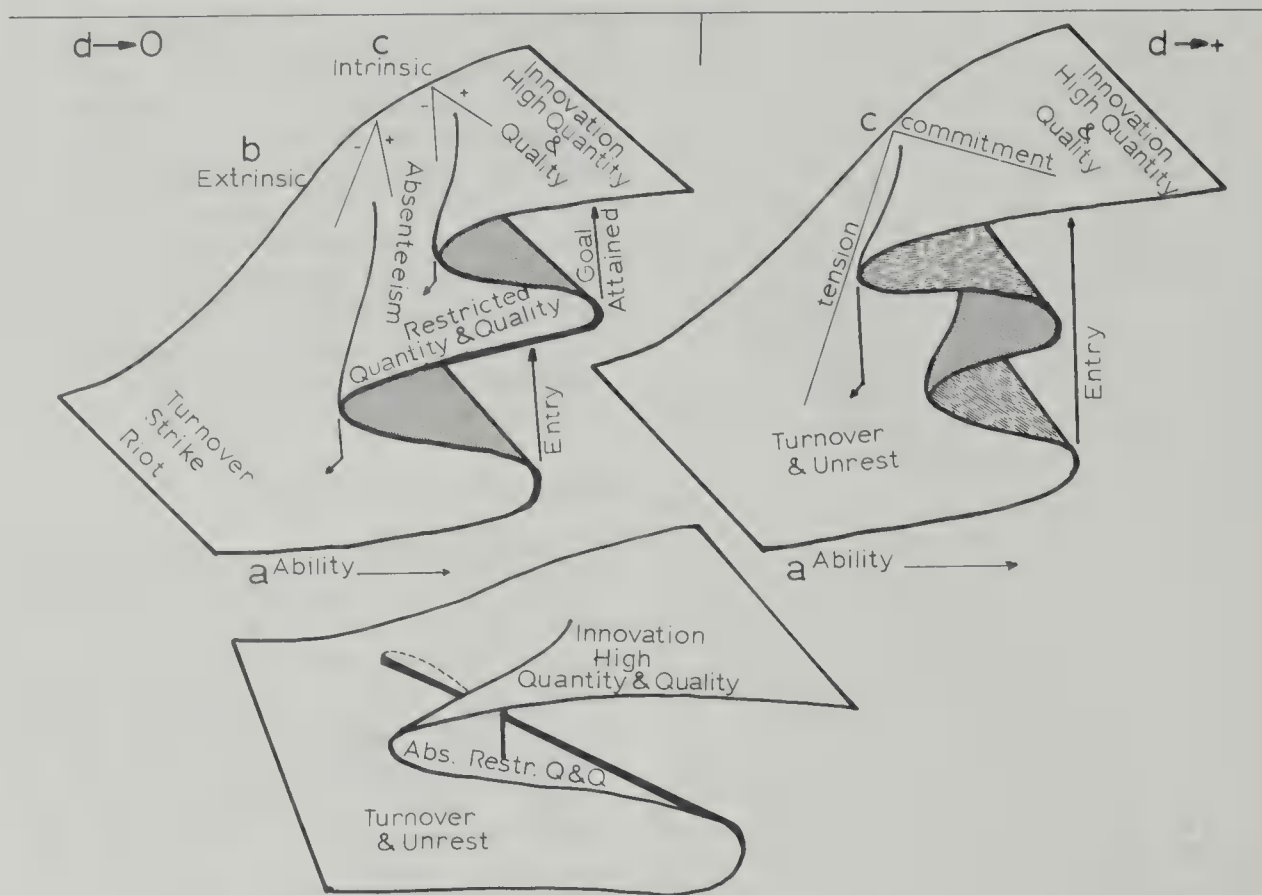


Figure 5. Three sectionings of the butterfly model of motivation in organizations. (The center sectioning is the most common display of the surface illustrating the pocket when c and d parameters are held constant. The left and right models allow c to vary given levels of d .)

the upper mode overall; there would be less disparity among subjects on this issue.

The lower mode would describe persons who leave the organization voluntarily or are fired for chronic absenteeism or poor performance. In organizations that do not have an organized absenteeism policy, chronic absentees would be noted at this level. Strikes and riots are expected in extreme conditions. Turnover is the asymmetric reverse of organizational entry (see Figure 5).

Controls. Control parameters in the hypothesized model are ability (a), extrinsic motivation (b), intrinsic motivation (c), and organizational climate (d). Ability and intrinsic and extrinsic motivation were mentioned already in conjunction with cusp and swallowtail submodels. Organizational climate appears for the first in the butterfly parameter position of the model. Ability is broadly defined and includes any relevant cognitive tests or work samples, or biodata items that might be used in a personnel selection scheme in place of ability. Ability is thus relative to the research questions; in predicting absenteeism, ability to attend from the Steers and Rhodes (1978) process model would contribute to the a parameter. Variables b and c denote a range of specific motivational influences. The classification of an item into the intrinsic or extrinsic category depends for the present purposes on whether its effect on behavior is more similar to the b or c control parameter. In addition, any demographic variable that is otherwise known as a moderator of ability could be a b variable.

Climate in the model consists of both subjective and objective variables, (Heller, Guastello, & Aderman, 1982; James & Jones,

1974; Jones & James, 1979) and describes differences among gross types of organizations (e.g., work, volunteer, academic), differences among organizations within type, and differences among subgroups within an organization (Drexler, 1977). Part of what is currently conceptualized as subjective climate would behave in the model as a c parameter, specifically those aspects that enhance or hinder intrinsic motivation. Parameter d governs the coaction of intrinsic and extrinsic motivation types (Deci, 1975; Enzle & Ross, 1978; Pritchard et al., 1977) plus exposures that vary by organization or subunit rather than by individual employee. Innovation climate (Siegel & Kaemmerer, 1978) and climate for achievement (deCharms, 1976; Litwin & Stringer, 1966) are subsets of a more general climate construct (Schneider & Reichers, 1983). Figure 5 shows a sectioning of the butterfly surface over parameter d . For low values of d , intrinsic and extrinsic motivation are additive, and a relatively greater number of behavior changes would occur between adjacent modalities. For higher values of d , intrinsic and extrinsic motivation are additive, and a relatively greater number of behavior changes would occur between extreme modalities. Because d is continuously defined, it would follow that the additivity or interactivity of intrinsic and extrinsic motivation is actually a matter of degree. Leadership style and organizational policies and practices would determine the specific relative contributions of intrinsic and extrinsic motivation to the behavior spectrum.

Gradients. The surface gradients of the butterfly surface describe changes in behavior from an ambiguous (unstable) state to one of stable modalities, often illustrating interesting theoret-

ical properties. Gradients in the motivation model describe satisfaction, equity, tension, and commitment constructs. The role of satisfaction in the opponent process was described earlier. Equity, tension, and commitment have been introduced in catastrophe models before, and the principles are summarized below.

Equity theory as proposed by Adams (1965) can be thought of as a statistical concept (Guastello, 1981). Although a person has a reasonably clear notion of what constitutes fair or unfair exchanges, some deviation is tolerated. This approximating process occurs in a single exchange, but more so in repeated exchanges. Equity judgements are made *relative* to something and involve a social comparison process. If a person's outcome to input ratio exceeds that of a comparison person or other reference point, a relative reward is experienced. If the ratio is less than the reference, a relative cost is experienced.

According to most theories, people process intrinsic and extrinsic rewards and costs differently. It is required, therefore, that they be separated in the model. In a job relationship the individual is judging four varieties of signal that comprise the surface gradients: intrinsic rewards, intrinsic costs, extrinsic rewards, and extrinsic costs. The gradients are rotational transformations of the control parameters described earlier. Individuals gravitate toward preferred types of exchanges, varying from low reward potential for low cost to high reward for high cost. Modal stabilities are stronger and behavioral changes more dramatic when the stakes are high (Guastello, 1981).

Tension and commitment were gradient-rotated control parameters in the cusp for turnover (Sheridan & Abelson, 1983) and represented the positive and negative aspects of intrinsic motivation. Tension arises from a job that is interesting, arousing, and novel in an unpleasant way. Commitment is the positive side and is a hybrid composed of intrinsic interest, job involvement, and motivating qualities of the individual and social environment. Steps within the Mobley (1977) turnover process can be identified as points along the cusp surface. Behavioral intention, the step closest to actual turnover, is anchored next to the fold line at which the behavior change occurs.

Application to Academic Performance

The application of the butterfly model of motivation theory to academic performance addresses a common problem. Consider unsuccessful freshmen where level of success is measured by grade point average (GPA). Students may perform poorly because of insufficient ability, poor study habits, negative attitude about grades, disinterest, or unfocused interest in the subject matter studies. Psychological adjustment problems may frequently be responsible for such an attitude. It is expected that the period of adjustment that affects nearly all students is generally completed by the end of the freshman year.

Hypothesized controls in the process are scholastic ability (*a*), interest in external features of a college education (*b*), and interest in the subject matter (*c*). Climate factor (*d*) was not varied in this application.

In addition to the theoretical objective of verifying that butterfly dynamics do occur in this situation, there are two practical objectives. The first is to describe the characteristics of successful and unsuccessful students, and persons of exceptional talent. The second is to enhance prediction in the neighborhood

of the critical college grade point average of 2.0, at which special administrative decisions must be made. The relative contributions of ability and interest motivation must be considered in counseling the failing student.

Method

Subjects

Subjects were 272 freshmen men and women from a midwestern technical university. It was standard procedure for all incoming freshmen to participate in the counseling program on a voluntary basis. Of these, 27% actually did participate and completed all the measures described below. Subjects were drawn from two consecutive classes of entering freshmen and were subdivided into a validation group of 191 and a cross-validation group of 81.

Measures

Records offered the following scores: American College Test composite score (ACT); realistic, investigative, and enterprising themes and academic orientation (AOR) from the Strong-Campbell Interest Inventory (SCII; Campbell, 1977); second semester freshmen grade point average (identified below as Time 2); and last year high school GPA (Time 1). High-school GPA was computed from grades from the senior year to a scale equivalent to the 4-point college GPA.

Modeling

Surface. The criterion for the freshmen problem is the difference between (a) high school senior GPA and (b) second term freshman GPA. The latter was preferred to cumulative GPA because the cumulative measure, inasmuch as it is a moving average, dampens the positive and negative fluctuations in performance. The criterion pattern is modeled as a butterfly. The lower mode is the case of a voluntary turnover during the semester, which was scored zero at Time 2, and those who would be terminated for not meeting the 2.0 standard. The middle mode represents those who perform within a small range around the class average, and the upper is the dean's list student, or those "most likely to make it into graduate school."

The GPA scales gives the illusion that high school students with a 4.0 can only show $\Delta y \leq 0$. In a catastrophe model that is not the case, because the criterion is transformed with regard to its standard deviation and lower limit. For a high-school-to-college sample, the lower limit can drop as much as 2.5 points, and variance would increase.

For instance, a college may select students with GPAs greater than or equal to 2.5, but once matriculated some would become failing students, due perhaps to simple inability, disinterest, incompatibility with the general atmosphere (climate) of the university, or other reasons. Hence the location parameter drops from 2.5 to 0.00. College GPA scores would be spread over a wider range and possibly polarized, hence the expected increase in variance. As a numerical illustration, let $\lambda_1 = 2.50$, $\sigma_1 = 0.50$, $\lambda_2 = 0.00$, and $\sigma_2 = 0.75$. Then a high school GPA of 4.00 becomes $z_1 = 3.00$, and a college 4.00 becomes $z_2 = 5.33$, hence $\Delta z = +2.33$, by maintaining a 4.00 GPA. If variance did not change, then $z_2 = 8.00$ and $\Delta z = +5.00$.

Ability. The American College Test composite score was adopted here for parameter *a* in the butterfly model.

Extrinsic motivation. The cost of education and financial aid are not analogous to pay in work situations. The financial aid system is structured so that the dollar cost to the student is relative to the student's resources. Any role cost would play is already sealed when the student finally chooses an institution to attend.

A closer analogy can be found, however, by separating motivation into two aspects: interest value of the subject matter and motivation to per-

Table 1
Summary of Regression for Butterfly Catastrophe and Control Analyses for Performance Differences

Predictor	F Model	F Beta	r	R	R ²
Butterfly difference					
Z ₁ ⁵		11.14***	-.48	.48	.24
Z ₁ ³		9.61**	-.54	.59	.35
Z ₁ ² *Investigative		0.29	-.56	.59	.35
Z ₁ ² *Realistic		0.21	-.54	.59	.35
Z ₁ ² *Enterprising		0.00	-.54	.59	.35
Z ₁ *Academic Orientation		0.13	-.47	.59	.35
ACT Composite	20.65***	1.80	.05	.59	.35
Control difference					
Realistic		4.54*	.11	.11	.01
Investigative		1.46	.00	.13	.02
Enterprising		0.01	.02	.13	.02
Academic Orientation		0.34	.00	.14	.02
ACT Composite	1.23	0.98	.06	.15	.02
Pre-post control					
High school GPA		19.84***	.26	.26	.07
Realistic		1.87	.04	.27	.07
Investigative		1.68	-.03	.28	.08
Enterprising		0.67	.06	.28	.08
Academic Orientation		0.38	-.01	.28	.08
ACT Composite	4.26***	2.27	.09	.30	.09

Note. ACT = American College Test; GPA = grade point average.
* *p* < .05. ** *p* < .01. *** *p* < .001.

form external features. The latter would include studying hard, working for grades, and selecting courses that will ensure a job or entrance to graduate or professional school. The distinction between two types of academic motivation are analogous to an “interesting job” compared to “anything that pays.” The external features of academic motivation are thus represented by SCII academic orientation as the *b* parameter.

Interest themes. SCII theme scores, and AOR as well, are generally lower at the beginning of the college career than later on. Patterns of interests crystallize over time. It would be expected that a freshman with highly developed interests would be intrinsically motivated to do well; a definite negative reaction might appear in a student of high motivation but low ability. It is also conceivable that low-interest persons would comprise the multimodal or high variance end of the relationship, in which case the high-interest theme subjects would show a unimodal performance-difference distribution. The specification of the polarities of theme scores is left for empirical analysis because they could be a function of unspecified organizational variables.

The realistic theme pertains to practical interests and to outdoor activities, with subscales for agriculture, nature, adventure, military, and mechanical activities. It is thought to be similar to interests of engineers. The investigative theme is an aggregate of medical and other science, mathematics, and service interests and would be characteristic of science majors and engineers. The enterprising theme represents interest in public speaking, law and politics, merchandising, sales, and business management; it would be most pertinent to business majors and engineers desiring preparation for engineering management. The realistic, investigative, and enterprising theme scores were all hypothesized to contribute to *c* parameter. Arts were virtually nonexistent, and social sciences served mainly a support role in the undergraduate curriculum at the school in question; thus, dominant aesthetic or social themes would not be of central concern, nor would the conventional theme. Because realistic, investigative, and enterprising themes all contribute

to intrinsic motivation in the setting in question, they are all hypothesized to contribute to the latent *c* control parameter.

Climate. Freshmen tend to take a variety of introductory courses; thus, their exposures are distributed among many of the departments of the institution by the end of the school year. In addition, they are exposed to certain commonalities among departments. With regard to climate, the school’s idiosyncrasies suggest that it is at greater variance with colleges in general than there is variance among subunits. For these reasons, *d* is replaced by a constant, 1.00.

The question that now arises is under what condition does the model builder substitute a constant for a control parameter rather than simply revise the hypothesis to a less complex model such as, perhaps, a swallowtail. The decision is based on the complexity of the behavior spectrum one expects to observe. If one expects three stable modes of performance, then a butterfly surface is required to account for such behavior, and the specification of four control parameters is required as well. If, on the other hand, only two out of three possible modes were expected with some unstable areas (e.g., students could drop out but no one did), then a model of lesser complexity such as a cusp or swallowtail would be appropriate.

The specific model in use for the academic performance application was, therefore,

$$\Delta z = \beta_0 + \beta_1 z_1^5 + \beta_2 1 z_1^3 + \beta_3 (\text{Realistic}) z^2 + \beta_4 (\text{Investigative}) z_1^2 + \beta_5 (\text{Enterprising}) z_1^2 + \beta_6 (\text{AOR}) z_1 + \beta_7 (\text{ACT}) + \beta_8 z_1^4, \quad (19)$$

where $z_1 = (y_1 - \lambda_1)/S_1 = (y_1 - 1.40)/.50$, and $z_2 = (y_2 - 0.00)/.75$. Values of location and scale were observed from the data before calculating the regression Equation 19.

Analyses

Primary. The butterfly difference Equation 19 was constructed and tested using forward selection regression procedure, with z_1^4 entered in last priority. Control equations corresponding to Equations 14 and 15 were also tested for comparison (Equations 20 and 21, respectively),

$$\Delta y = \beta_0 + \beta_1 (\text{Realistic}) + \beta_2 (\text{Investigative}) + \beta_3 (\text{Enterprising}) + \beta_4 (\text{AOR}) + \beta_5 (\text{ACT}); \quad (20)$$

$$y_2 = \beta_0 + \beta_1 (\text{High School GPA}) + \beta_2 (\text{Realistic}) + \beta_3 (\text{Investigative}) + \beta_4 (\text{Enterprising}) + \beta_5 (\text{AOR}) + \beta_6 (\text{ACT}). \quad (21)$$

Cross-validation and unit-weighting analyses were also performed on these three models. Subjects were assigned to the validation and cross-validation subsamples as follows: Subjects were ordered alphabetically in two sets, one for each class year of students sampled. They were serially assigned to subsamples, 7 to the validation group, followed by 3 to

Table 2
Summary of Cross-Validation and Unit-Weighting Analyses

Model	Validity (N = 191)	Cross-validity (N = 81)	Unit-weights (N = 272)
Butterfly-difference			
R ²	.34	.37	.23
F	13.58	46.97	81.66
Control difference			
R ²	.02	.02	.00
F	0.84	1.45	1.01
Pre-post control			
R ²	.10	.05	.02
F	3.48	4.10	4.63

the cross-validation group. No other heuristic was used to assign subjects.

Equation 19 was recalculated using the validation subsample of 181 cases; estimates of location and scale were based on the value obtained for all 272 cases. Empirical weights were obtained, and the predicted value of Δz was calculated for subjects in the holdout sample. Predicted and actual scores were correlated; again full-sample estimates of location and scale were used.

Unit-weighting (Wainer, 1978) was used to determine whether empirical weights or equal weights give more stable butterfly catastrophe regression results. Past statistical applications of catastrophe theory suggested that empirical weights were typically unequal in magnitude (Guastello, 1982b, 1982c; Guastello & McGee, 1982). Wherry's shrinkage formula was also applied to multiple regression equations.

Recoded. Utility transformation and IDFs (0–1 schemes described earlier) were constructed and tested in conjunction with the butterfly-difference equation. For the two-mode utility transformation, GPA scores below -1σ were recoded to 0.00; others were left unchanged (multiplied by 1.00). For the upper IDF, GPA scores less than 1σ were recoded to 0.00. For the lower IDF, GPA scores above -1σ were recoded to 0.00. After recoding scores, scores were normalized using the same estimates of location and scale that were obtained before recoding.

A special three-way utility transformation was constructed for the present problem. For GPA scores (both high school and colleges) of 3.5 or above, the normalized score was multiplied by 2.0. GPA scores less than 2.0 were recoded to zero. Scores between 2.0 and 3.5 remained normalized but otherwise unchanged (multiplied by 1.0). Thus, three distinct utility types were defined: the failing student, the dean's list student, and the average student. In other words, a trimodal distribution was created from the existing difference score distribution.

Results

Primary Analyses

The butterfly-difference model accounted for 35% of the second term freshman criterion variance ($p < .001$); only the power potentials accounted for unique portions of criterion variance. The simple correlations were all of about equal size, indicating considerable redundancy among the terms (see Table 1). The simple correlations were all negative, which for the interest variables meant that multimodality was associated with low theme interest scores. The quartic term was tested but not retained in the model due to insufficient tolerance (partial r to enter = .16, tolerance = .00012, $F = 7.10$), meaning that any contribution the variable could make would be within rounding error, in this instance. It was thus given a trivial weight in the model of 0.00.

By contrast, the control difference equation accounted for 2% of the criterion variance. The weight for the realistic theme score was significant overall. The conventional selection model Equation 21 accounted for only 9% of its criterion-variance ($p < .01$), with high school GPA as the only term with significant weight.

Cross-validation. Equations 19, 20, and 21 all held up to initial expectations on the cross-validation analysis. The butterfly model was found to be stable, and the cross-validation actually showed a 3% improvement over the validation sample. Such an improvement, however, is only interpreted as due to chance fluctuation. For all three models, cross-validation of empirical weights resulted in better prediction than was achieved by unit-weighting the predictors (see Table 2).

Table 3

Summary of Regression for Cusp and Butterfly Lower IDF and Three-Way Utility Code

Predictor	<i>F</i> Model	<i>F</i> Beta	<i>r</i>	<i>R</i>	<i>R</i> ²
Butterfly lower IDF					
Z ₁ ⁵		13.49**	-.75	.75	.57
Z ₁ ³		16.67**	-.80	.83	.69
Z ₁ ² *Realistic		0.23	-.81	.83	.69
Z ₁ ² *Investigative		0.00	-.81	.83	.69
Z ₁ ² *Enterprising		2.49*	-.80	.83	.69
Z ₁ *Academic Orientation		1.64	-.80	.83	.70
ACT Composite	86.01**	0.64	-.09	.83	.70
Butterfly with three-way utility code					
Z ₁ ⁵		10.25**	-.68	.68	.46
Z ₁ ³		9.92**	-.72	.75	.56
Z ₁ ² *Realistic		0.10	-.71	.75	.56
Z ₁ ² *Investigative		0.06	-.73	.75	.56
Z ₁ ² *Enterprising		0.76	-.73	.75	.56
Z ₁ *Academic Orientation		1.50	-.68	.75	.56
ACT Composite	49.50**	2.82	.06	.75	.57

Note. IDF = implicit discriminant functions; ACT = American College Test.

* $p < .10$. ** $p < .001$.

Recoded Analyses

The utility transformation and the upper and lower IDF butterflies improved prediction with R^2 coefficients of .47 ($F = 33.67$), .45 ($F = 31.26$), and .70 ($F = 86.01$), respectively. The lower IDF, which was the best of the four butterflies described up to this point, indicated that the zone of discontinuous change was located around -1σ , rather than in the center of the score distribution, meaning that the greatest discontinuity in performance occurred around the Time-2 GPA of 1.6. Equivalent numbers of students crossed this threshold in the positive and negative directions. Many more remained in the no-change area. A detail of the lower IDF appears in Table 3.

Scatterplots. Figure 6 is a plot of difference scores as a function of the lateral axis, which is the asymmetry term. Such diagrams when made for the cusp are two-dimensional projections of three-dimensional data, and represent two views: facing the surface, which shows the distance between modes, and top down, which shows the surface bifurcation (Guastello, 1982a, 1982b, 1982c). As might be expected, the cusp-like attributes are more pronounced for higher R^2 values.

Plots of criterion differences by asymmetry show the butterfly trimodality, but as two-dimensional projections of five-dimensional data. The surface shape is obscured to a greater extent for low values of R^2 , and trimodality is progressively more obvious as R^2 increases.

Three-way utility. An R^2 coefficient of .57 was obtained for the three-way utility coded criterion model. Simple correlations hovered around .70, with the exception of one for ACT scores. A detail of the analysis appears in Table 3. The quartic term was tested in last priority for two- and three-way models and IDFs, but in all instances it was retained in the model with a weight of 0.00.

The Wherry formula for shrinkage was applied to the pri-

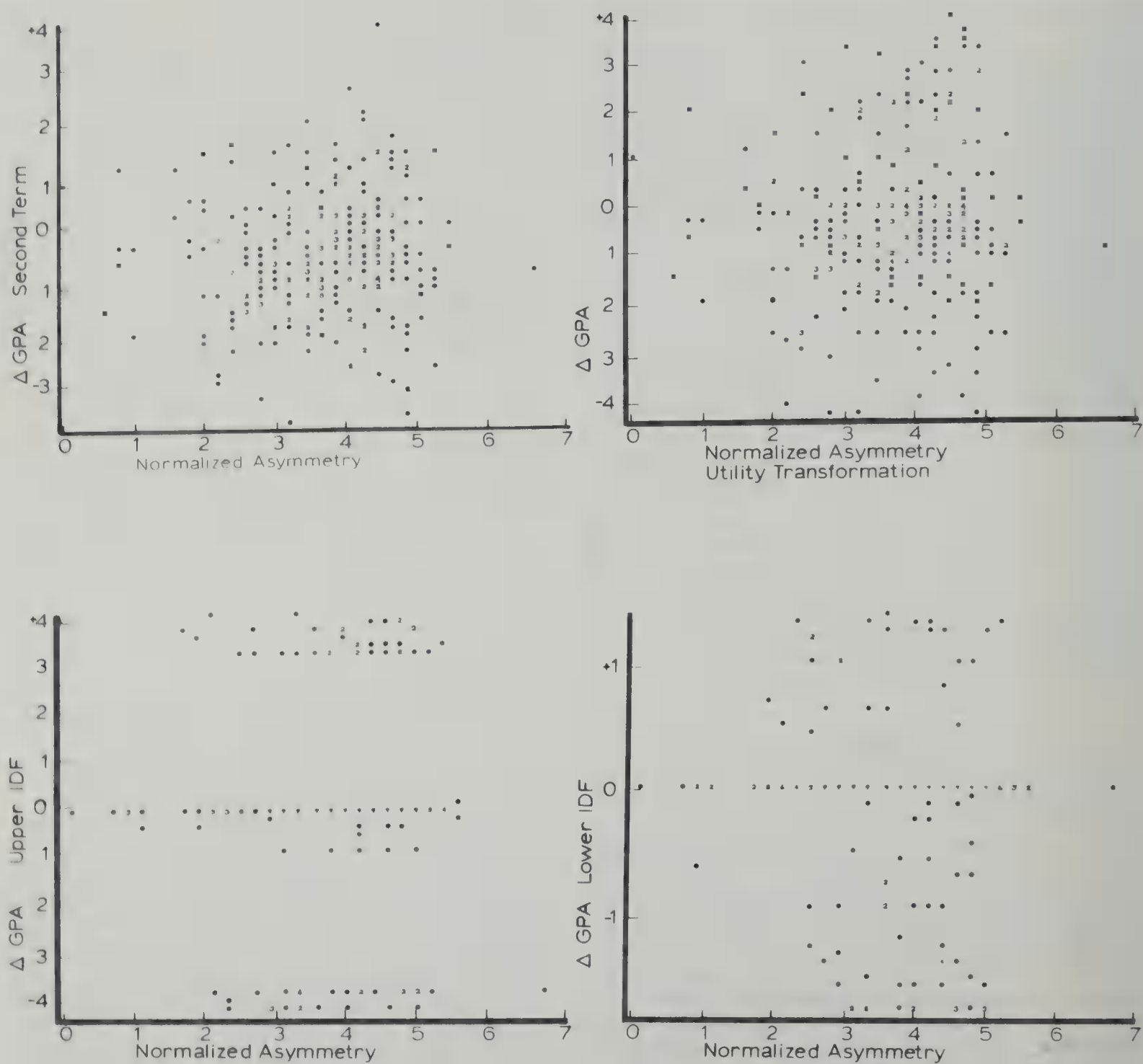


Figure 6. Scatterplots of data for uncoded, two-way utility, and implicit discriminant functions (IDF) analyses.

mary butterfly and all recoded schemes. In no case did R^2 drop more than .01.

Discussion

Academic Model Interpretation

Primary and supporting analyses indicated that the butterfly hypothesis was clearly superior to the linear alternatives by the best R^2 criterion. The quintic power potential that classifies the model as one of the elementary cuspsoids, here the butterfly, accounted for a unique portion of criterion variance after all other terms in the model were entered. Bifurcation and bias interac-

tions all showed large simple correlations with the criterion, and are interpreted as making small contributions to the overall model. Low theme scores were associated with positive changes in relative GPA, and high theme scores were associated with a drop in performance. Had they been noncontributing variables, simple correlations closer to .00 would have been attained, and the butterfly model would have lost accuracy on cross-validation. ACT scores had negligible impact on performance, possibly because subjects were selected in part on that basis.

The lower IDF indicated that most catastrophic activity was centered at about -1σ , which was a Time-2 GPA of 1.6. Students likely to cross into the acceptable performance zone were those with high AOR, realistic, enterprising, and investigative

theme scores. High scores on each of these variables statistically inferred negative Δz , which, because of the retrograde transformation, really meant a positive shift in behavior. Geometrically, larger performance variability was also denoted by high AOR and theme scores.

Students who were on the low end of GPA from the selected high school sample, and who were also low interest and low AOR, were likely to drop out. When the whole sample was viewed on a nontransformation basis, however (Table 1 data), it was the student with the high interest and high grades in high school who showed the greatest drop in grades, relatively speaking. The three cases just mentioned are examples of mediocre high school students waking up once they have found a medium suited to their interests, inept students, and interested students who, once in the institution of choice, see no further need to maintain an exemplary scholastic record. The latter group might have a tendency toward noncompetitiveness.

The three-way utility analysis yielded results quite comparable in interpretation to the primary analysis, but with an elevation in R^2 from .35 to .57. The results speak well for not only the catastrophe hypothesis, but also for the use of utility-coded criteria, as Novick (1980) proposed. By coding performance scores according to the utilities associated with them, a trimodal distribution was created. The resulting scores are accurately predicted with the butterfly model.

Were difference scores obtained throughout the analyses simply regression to the mean effects, that is, statistical artifacts? No, for two reasons. The types of student mentioned did gravitate toward a mean level of response, but it was possible to interpret their behavior patterns as distinct and real. In light of student behavior phenomena, then, they are asymptotically approaching the middle (of three) stable modes of performance. The second reason is that students shifted away from the mean as well as toward it. If the butterfly effect were a complete artifact, trimodality of difference scores would not be observed in Figure 6. Once again all these changes are predictable using the butterfly model and the regression equation associated with it. The primary analyses (no special codings used) provide an inkling of how much of the butterfly effect was a simple linear relationship: not more than 2/35 to 9/35, depending on whether one compares the butterfly to a linear difference or a simple Time 2 model.

Other studies (e.g., Burke, 1982) have determined stronger relationships between high school GPA and college performance than were obtained here. Such studies, however, are characterized by two design details that enhance the relationship. First, cumulative GPA is generally used as the college criterion, which has the effect of flattening any sudden changes in performance that may be operating; the stability of performance inferred by such a measure is artificial. Additional stability is introduced by averaging high school GPA over all 4 years rather than just senior year.

The generalizability of the study is limited in one particular respect. Because subjects were voluntary participants in a counseling program, the sample is likely to be biased in favor of persons who needed counseling there most. Such students would typically have low SCII theme scores. According to the model developed in the study, low theme scores indicated

greater amounts of instability. Thus, the sample is likely to contain a larger than usual proportion of unstable performers.

Discontinuous Change Processes in Psychology

Catastrophe theory is a general system theory for describing and predicting discontinuous changes of events. The general finding that its polynomial regression models account for more criterion variance than conventional models is largely attributed to the use of power potentials. The degree of the polynomial denotes a function of a special level of complexity. It also describes an autonomous process associated with the behavior change. Thus, catastrophe theory provides new types of hypotheses that can be proposed and tested concerning the structure of change and critical points. Although the elementary catastrophe models do not describe every possible type of change, research and theory have indicated strong advantages to their use for motivation and performance dynamics.

References

- Abelson, M. A. (1982). *Catastrophe theory model of the employee withdrawal process leading to job termination*. (Doctoral dissertation, Pennsylvania State University, 1981). *Dissertation Abstracts International*, 42, 3279A. (University Microfilms No. 81-29, 129)
- Adams, J. S. (1965). Inequity in social exchange. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 267-269). New York: Academic Press.
- Adelman, I., & Hihn, J. M. (1982). Politics in Latin America: A catastrophe theory model. *Journal of Conflict Resolution*, 26, 592-620.
- Agu, M. (1983). A method for identification of linear or nonlinear systems with the use of externally applied random force. *Journal of Applied Physics*, 54, 1193-1197.
- Arnol'd, V. I. (1974). Normal forms of functions in the neighborhoods of degenerate critical points. *Russian Mathematical Surveys*, 29, 10-50.
- Arnol'd, V. I. (1981). *Singularity theory: Selected papers*. Cambridge, England: Cambridge University Press.
- Ayers, T. J. (1981). Catastrophe theory and brightness judgment. *Perception and Psychophysics*, 29, 407.
- Baker, J. S., & Frey, P. W. (1980). A cusp catastrophe: Hysteresis, bimodality, and inaccessibility in rabbit eyelid conditioning. *Learning and Motivation*, 10, 520-535.
- Balasko, Y. (1978). The behavior of economic equilibria: A catastrophe theory approach. *Behavioral Science*, 23, 375-382.
- Benham, C. J., & Kozak, J. J. (1976). Denaturation: An example of a catastrophe. II: Two-state transitions. *Journal of Theoretical Biology*, 63, 125-149.
- Benham, C. J., & Kozak, J. J. (1978). Catastrophes in statistical biophysics. *Behavioral Science*, 27, 26-42.
- Bigelow, J. (1982). A catastrophe model of organizational change. *Behavioral Science*, 27, 26-42.
- Bosserman, R. W. (1982). The internal security subsystem. *Behavioral Science*, 27, 95-103.
- Brown, J. S. (1948). Gradients of approach and avoidance responses and their relation to motivation. *Journal of Comparative and Physiological Psychology*, 41, 450-465.
- Bruter, C. P. (1978). The theory of catastrophes: Some epistemological aspects. *Synthese*, 39, 293-316.
- Burke, M. J. (1982). A path analytic model of the direct and indirect effects of mathematical aptitude and academic orientation on high school and college performance. *Educational and Psychological Measurement*, 42, 545-550.

- Callahan, J. (1980). Bifurcation geometry of E_6 . *Mathematical Modeling*, 1, 283–309.
- Callahan, J. (1982a). Special bifurcations of the double cusp. *Proceedings of the London Mathematical Society*.
- Callahan, J. (1982b). A geometric model of anorexia and its treatment. *Behavioral Science*, 27, 140–154.
- Campbell, D. T. (1977). *Manual for the Strong-Campbell Interest Inventory*. Stanford, CA: Stanford University Press.
- Carhart, D. H. (1984). *An examination of the potential use of quantified catastrophe theory in management: A case study, or can management tell which straw will break the camel's back?* (Doctoral dissertation, George Washington University, 1984). *Dissertation Abstracts International*, 45, 1267B. (University Microfilms No. DA84-14, 998)
- Casti, J. (1982). Catastrophes, control, and the inevitability of spruce budworm outbreaks. *Ecological Modeling*, 14, 293–300.
- Chidley, J., Lewis, P., & Walker, P. (1978). The cusp catastrophe as a market planning aid. *Behavioral Science*, 23, 351–354.
- Chillingworth, D. R. J. (1976). *Differential topology with a view to applications*. London: Pitman.
- Cobb, L. (1978). Stochastic catastrophe models and multimodal distributions. *Behavioral Science*, 23, 360–374.
- Cobb, L. (1980). *Parameter estimation for the cusp catastrophe model: Programs and examples*. Charleston: Medical University of South Carolina.
- Cobb, L. (1981a). Parameter estimation for the cusp catastrophe model. *Behavioral Science*, 26, 75–78.
- Cobb, L. (1981b). Stochastic differential equations for the social sciences. In L. Cobb & R. M. Thrall (Eds.), *Mathematical frontiers of the social and policy sciences* (pp. 37–68). Boulder, CO: Westview Press & AAAS.
- Cobb, L. (1981c). Multimodal exponential families of statistical catastrophe theory. In C. Taille, G. P. Patil, & B. Baldessari (Eds.), *Statistical distributions in scientific work* (Vol. 3, pp. 1–24). Holland: Reidel.
- Cobb, L., Koppstein, P., & Chen, N. H. (1983). Estimation and moment recursion relationships for multimodal distributions of the exponential family. *Journal of the American Statistical Association*, 78, 124–130.
- Cobb, L., & Watson, B. (1980). Statistical catastrophe theory: An overview. *Mathematical Modeling*, 1, 311–317.
- Cobb, L., & Zacks, S. (1983, August). *Applications of catastrophe theory for statistical modelling in the biosciences*. Paper presented at the meeting of the Biometric Society, Toronto.
- Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decision* (2nd ed.). Urbana: University of Illinois Press.
- deCharms, R. (1976). *Enhancing motivation: Change in the classroom*. New York: Irvington.
- Deci, E. L. (1975). *Intrinsic motivation*. New York: Plenum.
- DeGreene, K. B. (1978). Force fields and emergent phenomena in socio-technical macrosystems: Theories and models. *Behavioral Science*, 23, 1–14.
- Dendrinos, D. S. (1979). Slums in capitalist urban settings: Some insights from catastrophe theory. *Geographica Polonica*, 42, 63–75.
- Dendrinos, D. S. (1980). *Catastrophe theory in urban and transport analysis*. Washington, DC: U.S. Department of Transportation (DOT/RSPA/DPB-25/80/20).
- Dendrinos, D. S., & Mullaly, H. (1983). Optimum control in nonlinear ecological dynamics of metropolitan areas. *Environment and Planning A*, 15, 543–550.
- Ditrich, J. E., & Carrell, M. R. (1979). Organizational equity perceptions, employees' job satisfaction, and department absence and turnover rates. *Organizational Behavior and Human Performance*, 24, 29–40.
- Dockens, W. S. III (1979). Induction/catastrophe theory: A behavioral ecological approach to cognition in human individuals. *Behavioral Science*, 24, 94–111.
- Dold, A., & Eckman, B. (Eds.). (1976). *Structural stability, the theory of catastrophes, and applications in the sciences*. New York: Springer-Verlag.
- Drexler, J. A., Jr. (1977). Organizational climate: Its homogeneity within organizations. *Journal of Applied Psychology*, 62, 38–42.
- Dyer, L., & Parker, D. F. (1975). Classifying outcomes in work motivation research: An examination of the intrinsic-extrinsic dichotomy. *Journal of Applied Psychology*, 60, 455–458.
- Edwards, J. E., & Waters, L. K. (1981). Moderating effect of achievement motivation and locus of control on the relationship between academic ability and academic performance. *Educational and Psychological Measurement*, 41, 585–587.
- Enzle, M. E., & Ross, J. M. (1978). Increasing and decreasing intrinsic interest with contingent rewards: A test of cognitive evaluation theory. *Journal of Experimental Social Psychology*, 14, 588–597.
- Evan, D., & Zeiss, A. M. (1984). Catastrophic theory: A topological reconceptualization of sexual orientation. *New Ideas in Psychology*, 2, 235–251.
- Fararo, T. J. (1978). An introduction to catastrophes. *Behavioral Science*, 23, 291–317.
- Fishbein, M. (1967). Attitude and prediction of behavior. In M. Fishbein (Ed.), *Readings in attitude theory and measurement* (pp. 477–492). New York: Wiley.
- Flay, B. R. (1978). Catastrophe theory in social psychology: Some applications to attitudes and social behavior. *Behavioral Science*, 23, 334–350.
- Frey, P. W., & Sears, R. J. (1978). A model of conditioning incorporating the Rescorla-Wagner associative axiom, a dynamic attention process, and a catastrophe rule. *Psychological Review*, 85, 321–340.
- Fung, K. K. (1980). Benefits and costs of confidential information: An application of systems theory and catastrophe theory. *Behavioral Science*, 25, 192–204.
- Furutani, N. (1976a). A new approach to traffic behavior. I: Modeling of "follow-defense" behavior. *International Journal of Man-Machine Studies*, 8, 597–615.
- Furutani, N. (1976b). A new approach to traffic behavior. II: Individual car and traffic flow. *International Journal of Man-Machine Studies*, 8, 731–742.
- Furutani, N. (1977). A new approach to traffic behavior. III: Steering behavior and the butterfly catastrophe. *International Journal of Man-Machine Studies*, 9, 233–254.
- German, I. (1984). *A disequilibrium model and catastrophe theory: The case of an oil shock*. (Doctoral dissertation, Purdue University, 1983). *Dissertation Abstracts International*, 44, 848A. (University Microfilms No. DA 84-00, 357)
- Gibson, C. G. (1979). *Singular points of smooth mappings*. London: Pitman.
- Gibson, C. G., Wirthmuller, K., du Plessis, A. A., & Looijenga, E. J. N. (1976). *Topological stability of smooth mappings*. New York: Springer-Verlag.
- Gilmore, R. (1981). *Catastrophe theory for scientists and engineers*. New York: Wiley.
- Guastello, S. J. (1981). Catastrophe modeling of equity in organizations. *Behavioral Science*, 26, 63–74.
- Guastello, S. J. (1982a, November). *Prison riots: An empirical application of catastrophe theory*. Paper presented at the meeting of the Illinois Psychological Association, Chicago.
- Guastello, S. J. (1982b). Color matching and shift work: An industrial application of the cusp-difference equation. *Behavioral Science*, 27, 131–139.
- Guastello, S. J. (1982c). Moderator regression and the cusp catastrophe: Application of two-stage personnel selection, training, therapy, and policy evaluation. *Behavioral Science*, 27, 259–272.
- Guastello, S. J. (1984a). *Catastrophe modeling of the accident process: Risk dispersions for ten industrial classes*. Manuscript submitted for publication.
- Guastello, S. J. (1984b). Cusp and butterfly catastrophe modeling for

- two opponent process models: Drug addiction and work performance. *Behavioral Science*, 29, 258–262.
- Guastello, S. J. (1984c). A catastrophe theory evaluation of a policy to control job absence. *Behavioral Science*, 29, 263–269.
- Guastello, S. J. (1984d, August). *Catastrophes, accidents, forecasting, and control*. Paper presented at the meeting of the American Psychological Association, Toronto.
- Guastello, S. J. (1985a). Euler buckling in a wheelborrow obstacle course: A catastrophe with complex lag. *Behavioral Science*, 30, 204–212.
- Guastello, S. J. (1985b). Color matching throughout the work week: An industrial application of the swallowtail difference equation. *Behavioral Science*, 30, 213–218.
- Guastello, S. J., & Dizadji, D. M. (1984, May). *Catastrophe modeling of the accident process: Systemic control of risk for open pit and underground mines*. Paper presented at the meeting of the Midwestern Psychological Association, Chicago.
- Guastello, S. J., & McGee, D. W. (1982, August). *Physical strength and work performance: Development of hiring standards and a catastrophe model of muscular fatigue*. Paper presented at the meeting of the American Psychological Association, Washington DC.
- Haag, G., & Dendrinis, D. S. (1983). Toward a stochastic dynamic theory of location: A nonlinear migration process. *Geographical Analysis*, 15, 269–286.
- Heller, R. M., Guastello, S. J., & Aderman, M. (1982). Convergent and discriminant validity of psychological and objective indices of organizational climate. *Psychological Reports*, 51, 183–195.
- Herzberg, F., Mausner, B., & Snyderman, D. (1959). *The motivation to work*. New York: Wiley.
- James, L. R., & Jones, A. P. (1974). Organizational climate: A review of theory and research. *Psychological Bulletin*, 81, 1096–1112.
- Jiobu, R. M., & Lundgren, T. D. (1978). Catastrophe theory: A quasi-quantitative methodology. *Sociological Methods and Research*, 7, 29–54.
- Johnson, A. C. (1982). *An application of catastrophe theory to the structure and operation of a board of directors of a nonprofit organization*. (Doctoral dissertation, The Wright Institute, Berkeley, 1982). *Dissertation Abstracts International*, 43, 1649B. (University Microfilms No. DA 82-20, 797).
- Jones, A. P., & James, L. R. (1979). Psychological climate: Dimensions and relationships of individual and aggregated work environment perceptions. *Organizational Behavior and Human Performance*, 23, 201–250.
- Kapur, J. N. (1981). Competition, games and catastrophes. In D. K. Sinha (Ed.), *Catastrophe theory and applications*, (pp. 80–86). New York: Halsted Press.
- Keown, R. (1980). Catastrophe theory and law. *Mathematical Modeling*, 1, 319–329.
- Kessler, R. C., & Greenberg, D. F. (1981). *Linear panel analysis*. New York: Academic Press.
- Labouvie, E. W. (1980). Measurement of individual differences in intra-individual changes. *Psychological Bulletin*, 88, 54–59.
- Landy, F. J. (1978). An opponent process theory of job satisfaction. *Journal of Applied Psychology*, 63, 533–547.
- Lawler, E. E. III, & Porter, L. W. (1967). The effect of performance on job satisfaction. *Industrial Relations*, 7, 20–28.
- Leon, F. R. (1981). The role of positive and negative outcomes in the causation of motivation forces. *Journal of Applied Psychology*, 66, 45–63.
- Litwin, G. H., & Stringer, A. (1966). *Motivation and organizational climate*. Harvard University, Graduate School of Business Administration, Division of Research.
- Locke, E. A. (1976). The nature and causes of job satisfaction. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 1297–1350). Chicago: Rand McNally.
- Locke, E. A., Mento, A. J., & Katcher, B. L. (1978). The interaction of ability and motivation in performance: An exploration of the meaning of moderators. *Personnel Psychology*, 31, 269–280.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Lu, Y. C. (1976). *Singularity theory and an introduction to catastrophe theory*. New York: Springer-Verlag.
- Lueger, R. J. (1982, May). *The effects of information accumulation and knowledge of patient history on judgments of backpain*. Paper presented at the meeting of the Midwestern Psychological Association, Chicago.
- Mawhinney, T. C. (1979). Intrinsic \times extrinsic work motivation: Perspective from behaviorism. *Organizational Behavior and Human Performance*, 24, 411–440.
- McGrath, J. E. (1976). Stress and behavior in organizations. In M. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 1031–1062). Chicago: Rand McNally.
- Merrill, S. J. (1980a). Mathematical models of humoral immune response. In T. A. Burton (Ed.), *Modeling and differential equations in biology* (pp. 13–49). New York: Marcel Dekker.
- Merrill, S. J. (1980b). Limit cycles in a model of B-cell stimulation. In T. A. Burton (Ed.), *Modeling and differential equations in biology* (pp. 225–237). New York: Marcel Dekker.
- Mobley, W. H. (1977). Intermediate linkages in the relationship between job satisfaction and employee turnover. *Journal of Applied Psychology*, 62, 237–240.
- Morrow, J. D. (1983). *A rational catastrophe theory of war*. (Doctoral dissertation, University of Rochester, 1982). *Dissertation Abstracts International*, 43, 2783A. (University Microfilms No. DA 83-02, 143)
- Novick, M. (1980). Statistics as psychometrics. *Psychometrika*, 45, 411–424.
- Oliva, T. A., & Capdevielle, C. M. (1977). Collective bargaining as a catastrophe model. *Academy of Management Proceedings*, 37, 117–181.
- Oliva, T. A., & Capdevielle, C. M. (1980). Sussman and Zahler: Throwing the baby out with the bath water. *Behavioral Science*, 25, 229–230.
- Oliva, T. A., Peters, M. H., & Murthy, H. S. K. (1981). A preliminary empirical test of a cusp catastrophe model in the social sciences. *Behavioral Science*, 26, 153–162.
- Poston, T., & Stewart, I. (1976). *Taylor expansions and catastrophes*. London: Pitman.
- Poston, T., & Stewart, I. (1978a). *Catastrophe theory and its applications*. London: Pitman.
- Poston, T., & Stewart, I. (1978b). Non-linear modeling of multistable perception. *Behavioral Science*, 23, 318–334.
- Pritchard, R. D., Campbell, K. M., & Campbell, D. J. (1977). The effects of extrinsic financial rewards on intrinsic motivation. *Journal of Applied Psychology*, 62, 9–15.
- Rambo, W. W., Chomiak, A. M., & Price, J. M. (1983). Consistency of performance under stable conditions of work. *Journal of Applied Psychology*, 68, 78–87.
- Saari, D. G. (1977). A qualitative model for the dynamics of cognitive processes. *Journal of Mathematical Psychology*, 15, 145–168.
- Sallam, A. A., & Dineley, J. L. (1983). Catastrophe theory as a tool for determining synchronous power system dynamic stability. *IEEE Transactions on Power Apparatus and Systems*, PAS-102, 622–623.
- Saunders, P. T. (1980). *An introduction to catastrophe theory*. New York: Cambridge University Press.
- Schneider, B., & Reichers, A. E. (1983). On the etiology of climates. *Personnel Psychology*, 36, 19–39.
- Schofield, N. (1977). The logic of catastrophe. *Human Ecology*, 5, 261–271.
- Schofield, N. (1980). Catastrophe theory and dynamic games. *Quantity and Quality*, 14, 519–545.
- Sheridan, J. E. (1980). Catastrophe model of employee turnover among

- hospital nursing staff. *Academy of Management Proceedings*, 40, 161-165.
- Sheridan, J. E. (1985). Catastrophe model of employee withdrawal leading to low job performance, high absenteeism and turnover during the first year of employment in an organization. *Academy of Management Journal*, 28, 88-109.
- Sheridan, J. E., & Abelson, M. A. (1983). Cusp catastrophe model of employee turnover. *Academy of Management Journal*, 26, 418-436.
- Siegel, S. M., & Kaemmerer, W. F. (1978). Measuring the perceived support for innovation in organizations. *Journal of Applied Psychology*, 63, 553-562.
- Sinha, D. K. (Ed.). (1981). *Catastrophe theory and applications*. New York: Halsted Press.
- Smith, J. Q. (1980). The prediction of prison riots. *British Journal of Mathematical and Statistical Psychology*, 33, 151-160.
- Smith, J. Q., Harrison, P. J., & Zeeman, E. C. (1981). Applicable catastrophe theory. III: The analysis of some discontinuous processes. In D. K. Sinha (Ed.), *Catastrophe theory and applications* (pp. 23-52). New York: Halsted Press.
- Smith, W. C. (1980). Catastrophe theory analysis of business activity. *Management Review*, 69, 26-28, 37-40.
- Smith, W. C. (1984). Forecasting buyer behavior: Using catastrophe theory to gain insight into customer purchase decisions. *Management Review*, 73(4), 48-53.
- Steers, R. M., & Rhodes, S. R. (1978). Major influences on employee attendance: A process model. *Journal of Applied Psychology*, 63, 391-407.
- Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1-50). New York: Wiley.
- Stewart, I. N. (1980). Catastrophe theory and equations of state: Conditions for a butterfly singularity. *Mathematical Proceedings of the Cambridge Philosophical Society*, 88, 429-449.
- Stewart, I. N. (1981a). Bifurcation and hysteresis varieties for the thermal-chainbranching model with a negative modal parameter. *Mathematical Proceedings of the Cambridge Philosophical Society*, 90, 127-139.
- Stewart, I. N. (1981b). Application of catastrophe theory to the physical sciences. *Physica*, 20, 245-305.
- Stewart, I. N. (1982). Catastrophe theory in physics. *Reports on Progress in Physics*, 45, 185-221.
- Stewart, I. N., & Peregoy, P. L. (1983). Catastrophe theory modeling in psychology. *Psychological Bulletin*, 94, 336-362.
- Sussman, H. J., & Zahler, R. S. (1978a). Catastrophe theory as applied to the social and biological sciences: A critique. *Synthese*, 37, 117-216.
- Sussman, H. J., & Zahler, R. S. (1978b). A critique of applied catastrophe theory in the behavioral sciences. *Behavioral Sciences*, 23, 383-389.
- Tanyi, G. E. (1982a). Energy and biological evolution. I: The equilibrium states of biochemical processes. *Bulletin of Mathematical Biology*, 44, 501-535.
- Tanyi, G. E. (1982b). Energy and biological evolution. II: The mathematical structure of equilibrium states. *Bulletin of Mathematical Biology*, 44, 537-547.
- Tanyi, G. E. (1982c). Energy and biological evolution. III: Theoretical ecology and macromolecular self-organization. *Bulletin of Mathematical Biology*, 44, 549-555.
- Tesser, A. (1980). When individual dispositions and social pressure conflict: A catastrophe. *Human Relations*, 33, 393-407.
- Thom, R. (1975). *Structural stability and morphogenesis*. New York: Benjamin-Addison-Wesley.
- Thompson, J. M. T. (1982). *Instabilities and catastrophes in science and engineering*. New York: Wiley.
- Vroom, V. H. (1964). *Work and motivation*. New York: Wiley.
- Wahba, M. A., & House, R. J. (1974). Expectancy theory in work motivation: Some logical and methodological issues. *Human Relations*, 27, 121-147.
- Wainer, H. (1978). On the sensitivity of regression and regressors. *Psychological Bulletin*, 85, 267-273.
- Wegner, D. M., & Vallacher, R. R. (1977). *Implicit psychology*. New York: Oxford University Press.
- Weintraub, E. R. (1983). Zeeman's unstable stock exchange. *Behavioral Science*, 28, 79-83.
- Weiss, D. J., Dawis, R. V., England, G. W., & Lofquist, L. H. (1967). *Manual for the Minnesota Satisfaction Questionnaire*. Minnesota studies in vocational rehabilitation. University of Minnesota.
- Wilson, A. G. (1981). *Catastrophe theory and bifurcations: Applications to urban and regional systems*. Berkeley: University of California Press.
- Woodcock, A. E. R. (1978). Landscapes of change: Catastrophe theory and biological processes. *Behavioral Science*, 23, 390-401.
- Woodcock, A. E. R. & Davis, M. (1978). *Catastrophe theory*. New York: Avon.
- Yelen, D. R. (1980). A catastrophe model for the effects of a response set on a discrimination task. *Perception and Psychophysics*, 28, 177-178.
- Youngblood, S. A., Mobley, W. H., & Meglino, B. M. (1983). A longitudinal analysis of the turnover process. *Journal of Applied Psychology*, 68, 507-516.
- Zeeman, E. C. (1974). On the unstable behavior of stock exchanges. *Journal of Mathematical Economics*, 1, 39-49.
- Zeeman, E. C. (1976). A mathematical model for conflicting judgments caused by stress applied to possible misestimation of speed caused by alcohol. *British Journal of Mathematical and Statistical Psychology*, 29, 19-32.
- Zeeman, E. C. (1977). *Catastrophe theory: Selected papers 1972-1977*. Reading, MA: Addison-Wesley.
- Zeeman, E. C., Hall, G., Harrison, P. J., Marriage, H., & Shapland, P. (1976). A model for institutional disturbance. *British Journal of Mathematical and Statistical Psychology*, 29, 66-80.

Received June 4, 1984

Revision received November 25, 1985 ■

Published quarterly
by the
American Psychological
Association

Marygrove College Library
Detroit, Michigan 48221
PLEASE DO NOT REMOVE

Journal of
Applied
Psychology

Editor

Robert M. Glan

Associate Editors

Irwin L. Goldstein

Frank J. Landy

The *Journal of Applied Psychology* is devoted primarily to original investigations that contribute new knowledge and understanding to any field of applied psychology except clinical psychology. The journal considers quantitative investigations of interest to psychologists doing research or working in such settings as universities, industry, government, urban affairs, police and correctional systems, health and educational institutions, transportation and defense systems, and consumer affairs. A theoretical or review article may be accepted if it represents a special contribution to an applied field.

Editor

Robert M. Guion, *Bowling Green State University*

Associate Editors

Irwin L. Goldstein, *University of Maryland*

Frank J. Landy, *Pennsylvania State University*

Consulting Editors

Lewis E. Albright, *deRecat & Associates, San Francisco, California*
Earl A. Alluisi, *Air Force Human Resources Laboratory, Brooks Air Force Base, Texas*

Kenneth M. Alvares, *Frito-Lay, Dallas, Texas*

Phipps Arabie, *University of Illinois*

William B. Askren, *Universal Energy Systems, Dayton, Ohio*

Kathryn M. Bartol, *University of Maryland*

Bernard M. Bass, *State University of New York, Binghamton*

Robert S. Billings, *Ohio State University*

Philip Bobko, *University of Kentucky*

C. Alan Boneau, *George Mason University*

Walter C. Borman, *Personnel Decisions Research Institute, Minneapolis, Minnesota*

Donald E. Broadbent, *University of Oxford, England*

Wayne F. Cascio, *University of Colorado, Denver*

Margaret M. Clifford, *University of Iowa*

H. Peter Dachler, *Hochschule St. Gallen für Wirts & Sozialwissen, St. Gallen, Switzerland*

Dan R. Dalton, *Indiana University*

Mark L. Davison, *University of Minnesota*

Robyn M. Dawes, *Carnegie-Mellon University*

Fritz Drasgow, *University of Illinois*

Beverly Dugan, *New York Telephone, New York, New York*

E. Ralph Dusek, *Advanced Resource Development Corporation, Columbia, Maryland*

James L. Farr, *Pennsylvania State University*

Jack M. Feldman, *University of Texas, Arlington*

Jeffrey H. Greenhaus, *Drexel University*

Tove Helland Hammer, *Cornell University*

William C. Howell, *Rice University*

Daniel R. Ilgen, *Michigan State University*

Andrew S. Imada, *University of Southern California*

Lawrence R. James, *Georgia Institute of Technology*

Stanislav V. Kasl, *Yale University*

James G. Kelly, *University of Illinois*

Gary P. Latham, *University of Washington*

Edwin A. Locke, *University of Maryland*

Robert P. Lowman, *Kansas State University*

Ben B. Morgan, Jr., *Old Dominion University*

Karlene H. Roberts, *University of California, Berkeley*

Paul R. Sackett, *University of Illinois, Chicago*

Steven L. Sauter, *NIOSH, Cincinnati*

Frank L. Schmidt, *University of Iowa*

Neal Schmitt, *Michigan State University*

Lyle F. Schoenfeldt, *Texas A&M University*

Stanley E. Seashore, *University of Michigan*

Kirk H. Smith, *Bowling Green State University*

Patricia Cain Smith, *Bowling Green State University*

Barry M. Staw, *University of California, Berkeley*

Mary L. Tenopir, *American Telephone & Telegraph Company, New York*

James R. Terborg, *University of Oregon*

Gary L. Wells, *University of Alberta*

Gary A. Yukl, *State University of New York, Albany*

Sheldon Zedeck, *University of California, Berkeley*

Manuscripts: Submit manuscripts in quadruplicate to the Editor, Robert Guion, Department of Psychology, Bowling Green State University, Bowling Green, OH 43403, according to instructions elsewhere in this journal (see the table of contents). APA and the editors assume no responsibility for statements and opinions advanced by contributors to *Journal of Applied Psychology*.

Change of Address: Send change of address notice and a recent mailing label to the attention of the Subscription Section, APA, 30 days prior to the actual change of address. APA will not replace undelivered copies resulting from address changes; journals will be forwarded only if subscribers notify the local post office in writing that they will guarantee second-class forwarding postage.

Reprints: Authors may order reprints of their articles from the printer when they receive proofs.

Single Issues, Back Issues, and Back Volumes: For information regarding single issues, back issues, or back volumes write to Order Department, APA, 1200 Seventeenth Street, N.W., Washington, DC 20036.

Microform Editions: For information regarding microform editions write to either of the following: University Microfilms, Ann Arbor, MI 48106; or Princeton Microfilms, Princeton, NJ 08540.

Copyright and Permission: Authors must secure from APA and the author of reproduced material written permission to reproduce an article in full or text of more than 500 words. APA normally grants permission contingent upon like permission of the author, inclusion of the APA copyright notice on the first page of reproduced material, and payment of a fee of \$20 per page. Permission from APA and fees are waived for authors who wish to reproduce a single table or figure provided the author's permission is obtained and full credit is given to APA as copyright holder and to the author through a complete citation. Permission and fees are waived for authors who wish to reproduce their own material for personal use; fees only are waived for authors who wish to use more than a single table or figure of their own material commercially. Permission and fees are waived for the photocopying of isolated articles for nonprofit classroom or library reserve use by instructors and educational institutions. Access services may use abstracts without the permission of APA or the author. Libraries are permitted to photocopy beyond the limits of U.S. copyright law; (a) post-1977 articles, provided the per-copy fee in the code for this journal (0021-9010/87/\$00.75) is paid through the Copyright Clearance Center, 27 Congress Street, Salem, MA 01970; (b) pre-1978 articles, provided the per-copy fee stated in the Publishers' Fee List is paid through the Copyright Clearance Center, 27 Congress Street, Salem, MA 01970. Address requests for reprint permission to the Permissions Office, APA, 1200 Seventeenth Street N.W., Washington, DC 20036.

APA Journal Staff: Susan Knapp, *Executive Editor*; Leslie A. Cameron, *Director, Journals Program*; W. Ralph Eubanks, *Manager, Journal Production*; Lois Czapiewski, *Production Editor*; Jodi Ashcraft, *Advertising Sales Manager*.

The *Journal of Applied Psychology* (ISSN 0021-9010) is published quarterly (beginning in February) in one volume per year by the American Psychological Association, Inc., 1400 North Uhle Street, Arlington, VA 22201. Subscriptions are available on a calendar year basis only (January through December). The 1987 rates follow: *Non-member Individual*: \$40 Domestic, \$43 Foreign, \$50 Air Mail. *Institutional*: \$82 Domestic, \$89 Foreign, \$96 Air Mail. *APA Member*: \$30. Printed in the U.S.A. Second-class postage paid at Arlington, VA, and at additional mailing offices. POSTMASTER: Send address changes to the *Journal of Applied Psychology*, 1400 North Uhle Street, Arlington, VA 22201.

Journal of
Applied Psychology

Copyright © 1987 by the American Psychological Association, Inc.

May 1987

Volume 72, Number 2

-
- 187 Multiple Spans in Transcription Typing
Timothy A. Salthouse and J. Scott Saults
- 197 Effects of Self- and Competitor Goals on Performance in an Interdependent Bargaining Task
Vandra L. Huber and Margaret A. Neale
- 204 Goal Importance, Self-Focus, and the Goal-Setting Process
John R. Hollenbeck and Charles R. Williams
- 212 Goal Commitment and the Goal-Setting Process: Problems, Prospects, and Proposals for Future Research
John R. Hollenbeck and Howard J. Klein
- 221 Nurse Turnover as Reasoned Action: Development of a Process Model
Perry H. Prestholdt, Irving M. Lane, and Robert C. Mathews
- 228 Power Relinquishment Versus Power Sharing: Theoretical Clarification and Empirical Comparison of Delegation and Participation
Carrie R. Leana
- 234 Judges' Mediation of Settlement Negotiations
James A. Wall, Jr. and Dale E. Rude
- 240 Cognitive Categorization and Quality of Performance Ratings
Michael K. Mount and Duane E. Thompson
- 247 Effects of Raters' Stress on the Dispersion and Favorability of Performance Ratings
Shanthi Srinivas and Stephan J. Motowidlo
- 252 The Systematic Distortion Hypothesis, Halo, and Accuracy: An Individual-Level Analysis
Steve W. J. Kozlowski and Michael P. Kirsch
- 262 Further Investigation of Common Knowledge Effects on Job Analysis Ratings
Angelo S. DeNisi, Edwin T. Cornelius III, and Allyn G. Blencoe
- 269 Job-Related Stress, Social Support, and Burnout Among Classroom Teachers
Daniel W. Russell, Elizabeth Altmaier, and Dawn Van Velzen
- 275 Relative Weight, Smoking, and Mental Health as Predictors of Sickness and Absence From Work
Katharine R. Parkes
- 287 Effects of Role Loss on Work-Related Attitudes
Judith A. Schlenker and Barbara A. Gutek
- 294 The Social Psychology of Eyewitness Accuracy: Misleading Questions and Communicator Expertise
Vicki L. Smith and Phoebe C. Ellsworth

(Contents continued on next page)

- 301 Perceptual and Preferential Discrimination Abilities in Taste Tests

Moshe M. Givon and Arie Goldman

- 307 Role of Efficacy Expectations in Predicting the Decision to Use Advanced Technologies: The Case of Computers

Thomas Hill, Nancy D. Smith, and Millard F. Mann

Short Notes

- 315 Increasing Voting Behavior by Asking People if They Expect to Vote

Anthony G. Greenwald, Catherine G. Carnot, Rebecca Beach, and Barbara Young

Monograph

- 319 Employee Stock Ownership and Employee Attitudes: A Test of Three Models

Katherine J. Klein

Other

- 314 Instructions to Authors

Multiple Spans in Transcription Typing

Timothy A. Salthouse and J. Scott Saults
University of Missouri—Columbia

Transcription typing has been postulated to consist of four components involving (a) input of chunks from the source text, (b) parsing of the chunks into discrete characters, (c) translation of the characters into movement specifications, and then (d) execution of those specifications in the form of keystroke responses. This multicomponent perspective on typing implies that it should be possible to identify distinct measures of anticipatory processing that correspond to the different processing components or spans. This prediction was tested, and largely confirmed, in three studies in which typists were administered a variety of experimental tasks to obtain span measures corresponding to the extent of anticipatory processing in different components of typing. As expected, the spans became progressively smaller as the hypothesized processing moved from input (with an average span of 8.1 characters) to execution (with an average span of only 1.4 characters).

In several recent articles, Salthouse (1984, 1985a, 1986) proposed that transcription typing involves the registration and coding of source text into easily remembered chunks by the input component, the partitioning or decomposition of the chunks into discrete characters by the parsing component, the conversion of characters into movement specifications by the translation component, and then those specifications implemented as keystrokes by the execution component. Although quite plausible arguments have been advanced for why each of these processing components might be necessary in transcription typing, there is still no direct evidence for their existence.

Salthouse (1985a) recently reported a study that attempted to investigate the existence of distinct processing components by using procedures to assess three different types of anticipatory processing, or spans. As expected, the three measures had considerably different magnitudes, with a measure termed *copy span* averaging 13.2 characters, a measure of eye-hand span 4.0 characters, and a measure of the sensitivity to constraining context only 1.8 characters. Although these results are consistent with a multicomponent model of transcription typing, it would have been more convincing to have had at least two independent measures of each postulated component. The present studies attempted to provide such evidence by obtaining measurements of five different types of processing spans during typing. The measures are described under the heading of the hypothesized component that they are presumed to be assessing.

Input Component

As previously noted, the input process is postulated to be responsible for the initial registration of the source material into

relatively familiar chunks. Two procedures were used to determine the amount of information available to the input process during transcription typing. One involved assessing the copy span in a manner similar to that described by Salthouse (1985a; i.e., by determining the average number of characters correctly typed after the unexpected disappearance of the display). However, unlike the procedure in previous study, predictability of the source text was minimized by using randomly arranged four-letter words as the stimulus material. (Previous research by Fendrick, 1937, Salthouse, 1985a, and West & Sabban, 1982, indicated that unrelated words are typed nearly as rapidly as meaningful text, and thus the use of this material was not expected to disrupt normal typing.)

The second measure of the information capacity of the input component was designated the *detection span*. It was assessed by determining how far in advance of the current keystroke the typist can detect a specially designated target character. If the target can be registered on the basis of features that do not require much processing, the number of characters intervening between the position of the target and the character currently being typed may serve as an index of when information first becomes available to the input process.

Parsing Component

In the parsing phase it is assumed that individual characters are isolated from the larger verbal units (words or phrases) of the source text, or from the chunks held in the input process. Two procedures were used to examine processing during the parsing component—one assumed to correspond to when the parsing began and the other postulated to reflect when it ended. Because the rate of typing is slowed if parsing does not proceed rapidly enough to ensure a continuous supply of information to later processes (cf. Salthouse, 1984), the quantity assessed by the eye-hand span can be interpreted as an indication of how far in advance of the keystroke the source material is decomposed for subsequent processing.

An indication of when this parsing processing is completed can be obtained by determining the point at which the typist is

This research was supported by National Institute on Aging Research Career Development Award 1K04 AG00146-01A1 and R01 AG04226-01A1 to the senior author.

Correspondence concerning this article should be addressed to Timothy A. Salthouse, who is now at the School of Psychology, Georgia Institute of Technology, Atlanta, Georgia 30332.

insensitive to further alterations in the stimulus display. That is, if the source text is altered while the subject is typing and he or she does not respond to the alteration, then it can be inferred that the source text is no longer necessary for the processing of the character undergoing the alteration. The point at which typists relinquish further monitoring of the source text was termed the *replacement span*.

The stimulus display in the replacement span task was arranged such that certain critical target letters were replaced by other letters at varying distances in advance of the current keystroke. The instructions were to type the replaced character, and hence, if the typist continuously monitored the source text, he or she should always type the second character and never the first character. On the other hand, if the source is monitored only for a limited period of time, and then no longer consulted, the typist should frequently type the first rather than the second character, particularly if the replacement occurs in close proximity to the current keystroke.

The replacement interval that resulted in a 50% probability of typing the first character was defined as the replacement span and was hypothesized to indicate how far in advance of the keystroke the typist became committed to a particular character. Because the commitment presumably occurs as information is transmitted out of the parsing process, the replacement span was assumed to provide an estimate of the degree of anticipatory processing between the parsing and translation components of the model.

Execution Component

Processing in the execution component is postulated to consist of the implementation of movement specifications transmitted from the translation component. The number of units in the execution component during normal typing can be assessed with a *stopping span* procedure introduced by Logan (1982), in which typists were instructed to immediately stop typing on the occurrence of a specially designated stop signal. Characters that continue to be typed after the occurrence of the stop signal can be interpreted as reflecting the contents of the execution component because these keystrokes are apparently no longer subject to interruption or modification.

Two independent studies were conducted, but because both involved similar procedures they are described together. However, one important procedural difference between the two studies was introduced based on an observation derived during the conduct of the first study. Because of the ease with which the skilled typists handled several of the special tasks, we suspected that they might have more reserve processing capacity than less skilled typists to devote to complying with the additional requirements while still performing normal typing. If this suspicion were correct, the span estimates for these individuals could be misleading because the values might simply reflect the ability to perform concurrent tasks while typing, and not the true sizes of the relevant component buffers. This excess capacity hypothesis was investigated by administering a simple reaction-time task, both alone and simultaneously with normal typing. If faster typists have more reserve capacity available for additional processing than do slower typists, they should exhibit less dis-

ruption of their reaction-time performance when switching from the single- to the dual-task conditions.

Typists from a wide range of age groups were used to increase the generality of the results and also to pursue an intriguing finding from an earlier study. Salthouse (1984) reported that older typists were able to maintain proficient levels of typing performance despite age-related declines in the efficiency of component processes, apparently because they relied on more extensive anticipation of forthcoming keystrokes than did young typists. Only the eye-hand span measure of anticipatory processing was available in the earlier study, however, and it is of interest to determine whether a similar compensatory effect is apparent with other measures of preparatory processing.

Method

Subjects

Study 1. Forty-five typists between 18 and 70 years old each received \$10 to participate in a single session of between 1.5 and 2 hr. All were experienced electric-typewriter touch typists with a mean of 10.0 hr of typing per week over the last 6 months, and a range of 0 to 35 hr. The mean number of months employed with at least 10 hr per week of typing was 68.3 months, with a range of 0 to 288 months.

Study 2. Forty electric-typewriter touch typists between 18 and 64 years old each received \$10 to participate in a single session of between 1.5 and 2 hr. They had a mean of 10.5 hr of typing per week over the last 6 months, with a range of 0 to 45 hr, and had been employed in positions involving at least 10 hr per week of typing, for an average of 101.5 months, with a range of 0 to 408 months. None had participated in the previous study.

Apparatus

All typing was performed on an Apple IIe microcomputer with a hardware clock to allow recording of keystroke intervals to a resolution of 10 ms. The keyboard arrangement on this computer is very similar to that of the popular IBM Selectric typewriter, and the typists generally reported that the feel was quite satisfactory.

Procedure

Study 1. Five different typing tasks were each performed twice in a counterbalanced order with a sixth task, the Digit Symbol Substitution subtest from the Wechsler Adult Intelligence Scale (Wechsler, 1958), administered between the first and second sequence of the five tasks. The purpose of the Digit Symbol test was to assess the representativeness of the subjects by allowing a comparison of age trends with previously published results.

Task 1, preceded by several minutes of practice to become familiar with the keyboard and typing in the manner requested, was normal typing from printed copy. The typing was to be performed as rapidly and accurately as possible, but the (carriage) return key was not to be pressed because the typed copy, which was visible on the display monitor, would automatically wrap around to the next line. Furthermore, no attempt was to be made to correct errors. The typing selections were paragraphs 6 and 2 (for the first and second administration of the task, respectively) from Form B of the Nelson-Denny Reading Test (Nelson & Denny, 1960). These passages (and paragraphs 4 and 7 used in Study 2), contained between 1,179 and 1,270 characters, including normal punctuation and capitalization.

Tasks 2 through 5 each involved the assessment of a different form of typing span and in order to maximize comparability, the same type of

stimulus material—randomly arranged four-letter words—was used in each task. The to-be-typed material was displayed on a single line of the video monitor and arranged such that each keystroke caused the displayed material to move one space to the left. No visible copy was produced in these tasks, and the impression from the perspective of the typist was of controlling the rate of a leftward-scrolling marquee. A practice phase was administered in each task to ensure that subjects fully understood the instructions.

Task 2 was designed to assess what was termed the typist's copy span. The following instructions were given to the subjects in this task:

Now you will be typing randomly arranged words that will be displayed on the monitor; you will not be able to see what you have typed. Type what appears on the screen in as normal a fashion as possible. At certain points the screen will go blank, and you should continue typing as much as you were sure was on the screen. After you can't remember any more of the material that was on the screen, press the '/' key to have the display reappear and resume typing. Remember to try to type as normally as possible.

A total of 35 characters were always visible on a line, and the display was blanked 10 times at random intervals ranging from 10 to 40 characters throughout the 500-character (100 4-letter words separated by spaces) passage.

Task 3 was designed to assess the typist's eye-hand span, with a procedure very similar to that used by Salthouse (1985a). Instructions were as follows:

Now the material will always be visible on the screen, but the number of characters will be systematically varied from 11 down to 1. You will probably find typing with only a few characters visible rather strange, but try to type as normally as possible. The number of visible characters will decrease and increase several times as you type, but always just try to type normally.

The passage in this task consisted of 1,000 characters (200 words), with the visible window changing by 2 characters with every 25 keystrokes, first decreasing from 11 to 1, and then increasing from 2 to 10, decreasing again from 11 to 1, and so forth.

The fourth task was used to obtain an estimate of the typist's detection span. Task instructions were as follows:

In this task a large number of characters will always be visible on the display, but occasionally a capital letter will appear. Whenever you notice a capital letter anywhere on the line you should press the '/' key as soon as you can and then resume typing. The capital letters should not be typed as capitals, but whenever you detect an upper-case letter you should press the '/' key. Always try to type as normally as possible.

A total of 20 targets appeared randomly with from 10 to 40 intervening characters throughout the 500-character (100-word) passage, and 35 characters were always visible on the line.

Task 5 was designed to measure the replacement span. The following instructions were used to introduce this task:

In this task the material will always be lower case, but on some occasions a letter will be changed in the display. You should ignore the original character when this happens and type the "corrected" version that appeared most recently on the display. Remember to try to type exactly what appears on the screen in as normal a fashion as possible.

A total of 40 target letters were replaced during the 625-character (125-word) passage in this task. The number of visible characters was always 35, but the critical letter was replaced either 3, 5, 7, 9, or 11 characters to the right of the left edge of the screen. The new letter string created by the replaced character was always a word, and each letter position

occurred equally often as the target (i.e., the first letter was replaced as often as the second, third, and fourth letters).

Study 2. Six tasks were performed twice, each in a counter-balanced order with the Wechsler Digit Symbol Substitution (Wechsler, 1958), subtest administered between the first and second sequence of tasks. Tasks 1 (normal typing) and 6 (detection span) were identical to those of Study 1, and the replacement span task (Task 5) differed only with respect to the range of replacement intervals and the length of the passage. The replacement intervals ranged from 1 to 7 rather than from 3 to 11, and the passage length was doubled from 625 to 1,250 characters (125 to 250 words). These changes were considered desirable because the results from Study 1 indicated that the replacement spans were generally less than 5 characters, but the previous procedure included only one interval less than 5 characters.

The second task was an auditory reaction-time task in which responses were made by pressing a foot pedal containing a micro-switch. A total of 30 tone signals were presented in a block, with intervals between signals ranging from 2 to 6 s.

Task 3 was a composite of Tasks 1 and 2 in that subjects typed from printed copy while responding to auditory signals with foot pedal responses. The typing task was stressed both by instructions and by delaying the introduction of the concurrent reaction-time task until subjects had typed for about 30 s.

Task 4 was designed to assess the typist's stopping span. The procedure and stimulus materials were identical to the copying span task of Study 1, except that the typist was instructed to stop typing as quickly as possible upon the disappearance of the display.

Results

Typing skill was assessed in terms of net typing speed, derived by subtracting five characters (one word) for each error from the gross number of keystrokes typed, dividing the net keystrokes by five to yield net words, and then dividing this quantity by the number of minutes required to type the entire passage. The mean net words per minute (wpm) across the 45 typists of Study 1 was 57.0, with a range from 23.5 to 85.3, and that for the 40 typists of Study 2 was 58.6, with a range from 18.4 to 111.5. Gross typing speeds uncorrected for errors, which were correlated .99 with net typing speeds in both studies, ranged from 25.4 to 91.7 wpm in Study 1, and from 21.9 to 122.6 wpm in Study 2, with means of 61.1 and 63.1, respectively. Because the net measure incorporates both speed and accuracy and is the measure most often used to represent typing skill outside the laboratory, all subsequent analyses are based on this measure.

The means and standard deviations across typists of the median interkey intervals (in ms) in the different conditions of the two studies are presented in Table 1. (The values for the eye-hand span task are not represented because the interkey interval varies systematically with the size of the preview window, which increased and decreased while the typist was performing in this task.) Typing rate was somewhat slower in each of the experimental conditions than in normal typing, but at least some of the rate reduction may be attributable to the frequent pauses mandated by the requirements to make special detection or restart responses.

Separate span measures were derived for each typist in both administrations of every task. The median number of characters typed after the disappearance of the display served as the estimate of the span in both the copy span and stopping span tasks.

Table 1
*Means and Standard Deviations Across Subjects
of Median Interkey Intervals*

Interval	Study 1		Study 2	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Normal typing	182	52	181	64
Concurrent typing	—	—	185	62
Detection span	223	61	217	66
Copy span	206	61	—	—
Replacement span	209	53	203	63
Stopping span	—	—	201	65
Reaction time				
Alone	—	—	269	49
Concurrent	—	—	431	85

The detection span was defined as the median number of characters intervening between the target and the character currently being typed. In other words, if on three separate occasions the target was detected when it was 12, 9, and 8 characters in advance of the leftmost character on the display, the detection span would be 9.

The eye–hand span was determined by first computing the median interkey interval at each preview window from 1 to 11, and then defining the eye–hand span as the largest window at which the median interval for that and all smaller windows was greater than the largest median across window sizes of 9, 10, and 11 characters. A somewhat different procedure was used in earlier studies (e.g., Salthouse, 1984, 1985a), but the present procedure was a convenient means of obtaining estimates from a single administration of the task, and informal comparisons indicated that it yielded values comparable to those derived from the previous procedures.

Replacement span was determined by computing the percentage of replaced (second) characters typed at each replacement interval, and then designating the span as the interval corresponding to a .5 probability of typing the replaced character.

Figure 1 portrays the frequency of typists with spans of each magnitude in Studies 1 and 2. Notice that the distributions are distinct and ordered in the manner one would expect if the spans reflected the operation of components becoming progressively more committed to a specific keystroke. In particular, note that the copy span was larger than the eye–hand span (a trend evident in 87% of the subjects for whom both measures were available), and the replacement span was larger than the stopping span (evident in all of the subjects for whom both measures were available). Detection span was expected to be larger than the replacement span but comparable to the copy span, and the results were generally consistent, as 80% of the typists had larger detection spans than replacement spans, whereas only 64% had detection spans larger than copy spans.

An alternative means of portraying the relations among the various anticipation measures is illustrated in Figure 2. The functions in this figure are based on data collapsed across all typists in a given study, and with the exception of the eye–hand span data that did not lend themselves to a representation in terms of probability, each represents a different type of probability. In the detection task the probability is that of detecting

the target by the indicated character position. Data points from right to left correspond to the cumulative probability that a target at that position, or one to the right of that position, will be detected. The functions for the copy span and stopping span tasks are also based on cumulative probabilities, in this case the probability that the typing will continue to at least the indicated character position. Data points for the replacement span task correspond to the probability that the original character will be typed instead of the replaced character.

An advantage of expressing the span results in a format like Figure 2 is that it represents the different anticipation measures along common axes and thus facilitates comparison of relative magnitudes. That is, this figure makes it clear that there are distinct differences between the probability of continuing to type with instructions to type as much as possible and the probability of continuing to type with instructions to stop as quickly as possible, and between both of those probabilities and the probability of detecting and responding to a change in the stimulus display. Moreover, the mean eye–hand span displayed in Figure 2 suggests that it is intermediate between the replacement span on the one hand, and the copy and detection spans on the other hand.

Several performance measures were obtained in the reaction-time tasks, beginning with the median of the 30 reaction times in each administration in the single- and dual-task conditions. The means of these medians are displayed in Table 1, along with the means of the median interkey intervals in normal typing alone, and normal typing with the concurrent reaction-time task. Typing errors increased from 1.6% to 2.2% from the single- to the dual-task conditions, but net typing speed only changed from 58.6 to 53.8 wpm. Moreover, typists at all skill levels exhibited comparable effects on typing performance because the correlation between typing skill (in net wpm) and ratio of typing performance (dual/single) was only .12. Because the greatest effect of the requirement to divide one’s attention was clearly on the reaction-time (RT) measures, subsequent analyses were confined to variables derived from these measures. In particular, absolute—RT(Dual)–RT(Single)—and relative—RT(Dual)/RT(Single)—indices of dual-task impairment were computed.

Correlation matrices for the relevant dependent variables in Studies 1 and 2 are presented in Tables 2 and 3, respectively. Values in parentheses along the diagonal are estimated reliabilities derived by applying the Spearman-Brown formula to the correlation between the values obtained in the two separate administrations of the task. All remaining correlations involving measures with more than one estimate are based on the average of the estimates from the two administrations of the task.

Discussion

Age Effects

The age correlations illustrated in Tables 2 and 3 indicate that age effects were relatively small on the various measures of anticipatory processing. One notable exception to this trend is the significantly positive correlation between age and replacement span in Study 1 (Table 1), suggesting that older typists commit to a stimulus character earlier than do young typists.

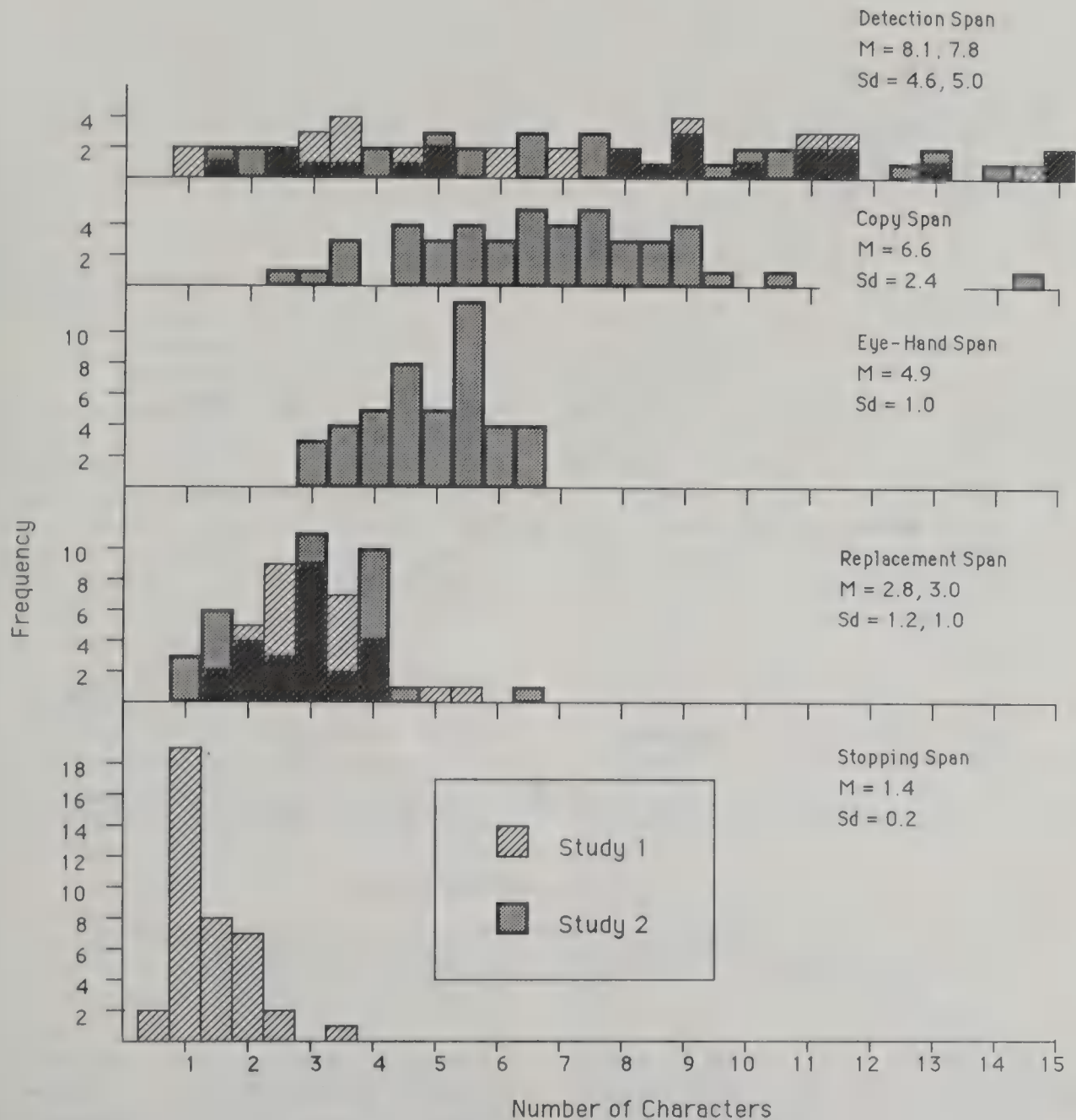


Figure 1. Distribution of subjects with each magnitude of span across the five span tasks in Studies 1 and 2.

Several factors might be responsible for the small age effects in the current studies relative to those found by Salthouse (1984), but unrepresentative sampling is apparently not one of them. The subjects in the present studies exhibited correlations between age and digit symbol score of $-.55$ and $-.42$, respectively, which are comparable to the values reported in several large-scale studies (e.g., see Salthouse, 1985b, for a review).

More plausible reasons for the smaller age effects are a restricted range of typing experience in Study 1 and underrepresentation of older ages in Study 2. That is, the typists in the Salthouse (1984) studies had maximums of 552 and 600 months of experience performing typing 10 hr per week or more, whereas the typists in the present Study 1 had a maximum of only 288 months. Furthermore, the correlation between age and this experience variable was above $.5$ in both of the studies in Salthouse (1984), but was only $.13$ in the present

Study 1. If the compensatory effects associated with aging require extensive experience for their development, it is probably unrealistic to expect them to be evident in samples that have only moderate amounts of experience and in which the older typists don't have considerably more experience than the young typists.

A relatively small percentage of older typists may account for the attenuated age effects in Study 2 because the oldest typist was only 64 years of age, compared to 68 and 72 years in the Salthouse (1984) studies. Perhaps even more important is that only 10% of the subjects in the current Study 2 were 60 years of age or over, whereas 18% of the subjects in both of the previous studies were in this age range.

One intriguing exception to the generally small age effects are the negative correlations between age and the two variables reflecting amount of dual-task interference. Older typists expe-

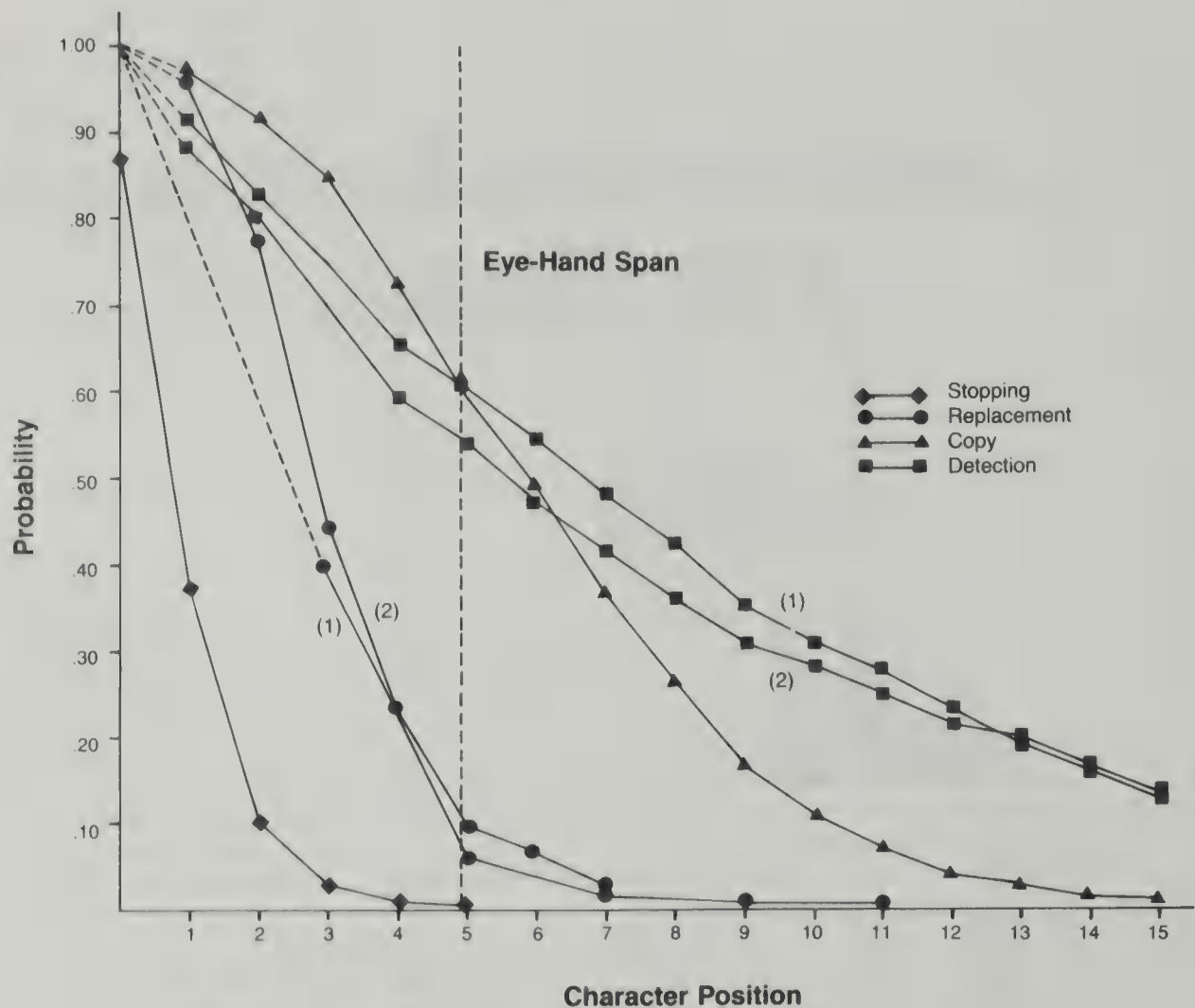


Figure 2. Probability functions across character positions for the different span tasks in Studies 1 and 2. (The dashed vertical line represents the mean eye–hand span across subjects.)

rienced more interference with a concurrent task while typing than did young typists, even though they were indistinguishable in overall level of typing proficiency when typing was performed alone. One interpretation of this result is that the older typists, relative to young typists, are using more of their available processing capacity performing the typing task and thus have less in reserve for the performance of additional simultaneous tasks. It therefore seems reasonable to infer that although the young and older typists perform at equivalent levels on the primary task, the older typists may be closer to their limits than are the young typists.

A similar finding that older adults experience greater abso-

lute increments in reaction time when performing a concurrent task than do young adults has been reported by Salthouse and Somberg (1982) and Somberg and Salthouse (1982). The present results extend the earlier ones, however, by demonstrating that the phenomenon is evident in both absolute and relative measures of dual-task interference, and is apparent even when the subjects are very experienced in the primary activity.

Skill Effects

Typing skill, as indexed by net words per minute, had moderately positive (and significant, $p < .05$) correlations with each

Table 2
Correlation Matrices for Study 1

Measure	1	2	3	4	5	6
1. Skill	(.98)	.01	.57	.15	.47	.46
2. Age		—	.08	–.33	.16	.42
3. Copy span			(.85)	.03	.39	.41
4. Detection span				(.91)	.14	–.12
5. Eye–hand span					(.54)	.38
6. Replacement span						(.71)

Table 3
Correlation Matrices for Study 2

Measure	1	2	3	4	5	6	7
1. Skill	(.98)	.07	.07	.80	.56	-.34	-.38
2. Age		—	-.07	.16	.13	.53	.35
3. Detection span			(.78)	-.15	-.26	.08	.14
4. Replacement span				(.84)	.71	-.08	-.11
5. Stopping span					(.83)	-.03	-.13
6. Reaction-time difference						—	.90
7. Reaction-time ratio							—

of the span measures except detection span. Faster typists, therefore, had larger copy spans, eye-hand spans, replacement spans, and stopping spans than did slower typists. Positive correlations have been reported previously for each span measure except replacement span (e.g., Logan, 1983; Salthouse, 1984, 1985a), although in several cases the earlier correlations were not significantly different from zero. These results clearly suggest that an important concomitant of typing skill is greater anticipation at each of several processing components.

Skill was negatively correlated with both measures of dual-task interference, indicating that faster typists experienced less interference in reaction time when that task was performed concurrently with typing. This finding is consistent with earlier reports (e.g., Dvorak, Merrick, Dealey, & Ford, 1936; Shaffer, 1975) that highly skilled typists are able to perform other activities while maintaining competent typing, but it also extends those reports by demonstrating that this effect systematically increases with increased skill.

Although it was hypothesized that more skilled typists might have larger spans in part because of their greater reserve capacity, the correlations between the measures of dual-task interference and estimates of detection span, replacement span, and stopping span were uniformly low. Unfortunately, because the reaction-time tasks substituted for the copy span and eye-hand span tasks in Study 2, the relation between dual-task interference and copy span and eye-hand span could not be determined. However, the available results are not consistent with the hypothesis that skilled typists had larger spans simply because they had greater surplus capacity to devote to the special requirements of the span tasks, because the correlations between the index of surplus capacity and span magnitude were quite small.

Multiple Spans

A major finding of Studies 1 and 2 is that measures of anticipatory processing thought to correspond to separate components of typing are of different magnitudes. These results are clearly consistent with the four-component model outlined earlier, and seem to necessitate a distinction among at least three phases of processing in any comprehensive theory of transcription typing. The correlational evidence was much more equivocal because although the reliabilities were generally moderately high, only the correlation between stopping span and replacement span in Study 2 was of comparable magnitude and these two spans were postulated to originate in different processing

components. Possible reasons for the low correlations among measures thought to reflect the same processing component are mentioned in the following discussion of specific spans.

The magnitude of the detection span was quite variable across subjects, possibly because several different strategies could be used in this task. On the one hand, subjects could simply try to type normally and emit a detection response only when a target was accidentally encountered near the occurrence of one's current keystroke. On the other hand, the subject could periodically decide to interrupt his or her typing to scan for targets, thus detecting the target at very great distances and resulting in larger detection spans.

The moderately high reliability coefficient suggests that the subjects were consistent in their utilization of a given strategy across the two administrations of the task, but it is unclear whether the measures represent the same concept in different subjects. Moreover, to the extent that the detection span does not represent a single entity; it may be unrealistic to expect it to be related to typing skill or to other span measures.

Copy span averaged 6.6 characters in Study 1, which is considerably smaller than the value of 13.2 obtained by Salthouse (1985a) and the value of 40 reported by Rothkopf (1980). Variations in task demands and stimulus material are probably responsible for most of these differences. Rothkopf required subjects to look at the source text and then try to type as much as they could remember, thus emphasizing memory rather than naturalistic typing. Salthouse (1985a) used meaningful text as the stimulus material, and the larger spans in that study may be a consequence of the easier predictability from prior context compared to the current study in which unrelated words were used as the source text.

The copy span is interpreted as indicating that average typists have about 6 to 7 characters in a temporary input buffer. The contents of this buffer are assumed to be easily coded chunks from the source text, and thus would be expected to vary in size with the familiarity and redundancy of the material.

The eye-hand span in Study 1 averaged 4.9 characters, which is somewhat higher than the estimates of 3.4 to 4.0 reported in earlier studies (e.g., Salthouse, 1984, 1985a). Inspection of the distribution of spans revealed that, compared to the previous studies, the current study had fewer subjects with spans below 4.0 but a very similar maximum span. The fact that the current study used four-letter words as stimulus material, whereas the other studies all used meaningful text containing words of variable lengths, may contribute to this higher minimum, but the exact mechanism responsible is still unclear.

By definition, the eye-hand span refers to the number of characters necessary to ensure a normal rate of typing. Because the values from this indirect procedure are similar to those reported by Butsch (1932) in direct measurements of the focus point of the eyes relative to the character being typed, it seems reasonable to infer that the eye-hand span corresponds to important aspects of processing. According to the model proposed by Salthouse (1984, 1986), the eye-hand span originates in the parsing component as individual characters are isolated and then sent to the translation and execution components for further processing.

The replacement span estimates were very similar in Studies 1 and 2, despite much greater resolution with the procedure used in Study 2. The results indicate that average typists tend to commit to a particular character about three characters in advance of the keystroke for that character. If the stimulus is changed before this time, the change is detected and the second character is typed, but if the switch occurs later it is typically unnoticed and the initial character is typed. This pattern suggests that, on the average, subjects relinquish further monitoring of the source text about three characters in advance of the keystroke.

The final anticipation measure is the stopping span, which averaged about 1.8 characters. This value is consistent with those reported by Logan (1982) in several variants of the stopping span task, and is also of the same magnitude as error detection responses observed when subjects are required to indicate when they have made an error (e.g., Long, 1976; Rabbitt, 1978; Shaffer & Hardwick, 1969). These latter findings are relevant because error detection can be assumed to function like a stop signal, and the responses to that signal typically occur within one to two keystrokes. Also in this range are the estimates of the amount of prior context found to influence the variability of a given keystroke (Gentner, 1982, 1983; Salthouse, 1985a). All of these phenomena are postulated to reflect the contents of the execution buffer containing already translated movement specifications no longer under the control of the subject.

Notice that although both the replacement span and the stopping span are interpreted as indices of points of commitment on the part of the typist, they are postulated to differ in the types of commitment involved. The replacement span indicates that the subject is committed to typing a particular character *if a keystroke response is to be made*, whereas the stopping span represents the number of responses that the subject is *committed to making*. In terms of the model outlined earlier, the replacement span corresponds to the characters sent from the parsing component to the translation and execution components, whereas the stopping span corresponds to the number of translated response codes in the execution buffer.

This distinction between the two types of commitment is on rather tenuous grounds because although all of the subjects had larger replacement spans than stopping spans, the largest cross-task correlation in Tables 2 and 3 was between the stopping span and the replacement span, implying that they may involve the same process. In fact, Logan (personal communication, November 1985) suggested that the two spans may reflect the same type of commitment and yield different estimates primarily because the stimuli in the two tasks differ in saliency.

It therefore seemed desirable to conduct a supplemental

study specifically focusing on possible differences and similarities between the stopping span and the replacement span in order to clarify this potentially important distinction between commitment to a character given that a keystroke is to be executed, and commitment to execute a keystroke. The approach used to investigate possible similarities and differences between the stopping span and replacement span tasks involved a hybrid task in which typists were instructed to stop typing whenever they detected a stimulus replacement. That is, as in the replacement task the typists were to type the most recent material that appeared on the display, but now whenever they noticed a character replacement they were to stop typing for several seconds before resuming their normal typing.

An additional task assessed the probability of detecting, and rapidly responding to, a stimulus replacement when no typing was required. The stimulus display consisted of right-to-left scrolling of the text as in the typing tasks, but the subject was instructed not to type and to press a key ('/') as rapidly as possible when a character substitution occurred in the display.

These new tasks, in addition to normal typing, the stopping span task, and the eye-hand span task, were administered to 10 typists ranging in age from 21 to 49 years and ranging in skill from 39.0 to 72.9 gross wpm.

Figure 3 portrays the results from this supplemental study in the same format as that used in Figure 2 to summarize the results of Studies 1 and 2. Comparison of the two figures reveals that the findings concerning the relative magnitudes of the stopping span, replacement span, and eye-hand span are replicated. Of particular interest is the finding that the functions for the replacement span and stopping span tasks are distinct even when the stopping response is triggered by the replacement stimulus. This suggests that although stimulus saliency may play a role in the stopping span task, it cannot account for all of the differences between the estimates derived from the stopping span and replacement span tasks.

Figure 3 also contains data corresponding to the probability of responding to (i.e., detecting) the character replacement in the replacement reaction time task and the stop-to-replace task. The function for the reaction-time task is rather noisy, perhaps because there was no constraint on eye position when the text scrolled without any typing, but it is still clear that detection probability was low when the replacement occurred near the left edge of the display. This trend is much more pronounced in the data from the stop-to-replace task, where the probability of detecting a stimulus change is greatest when it occurs three or four characters before the keystroke, but declines markedly as the replacement occurs earlier or later than this position.

These data on the probability of detecting the character replacement suggest that the stopping and replacement spans do not differ only in terms of the saliency of the stimulus because many of the replacement stimuli are not even detected. If the time to respond to the stimulus was the only factor responsible for replacement spans exceeding stopping spans, then one would expect the detection probabilities to be uniformly high, or at least to decline only at character positions greater than four. The results in Figure 3 indicate that this was not the case, and instead there was a systematic drop in probability of detecting a stimulus replacement as it occurred closer to the time of the keystroke for that character.

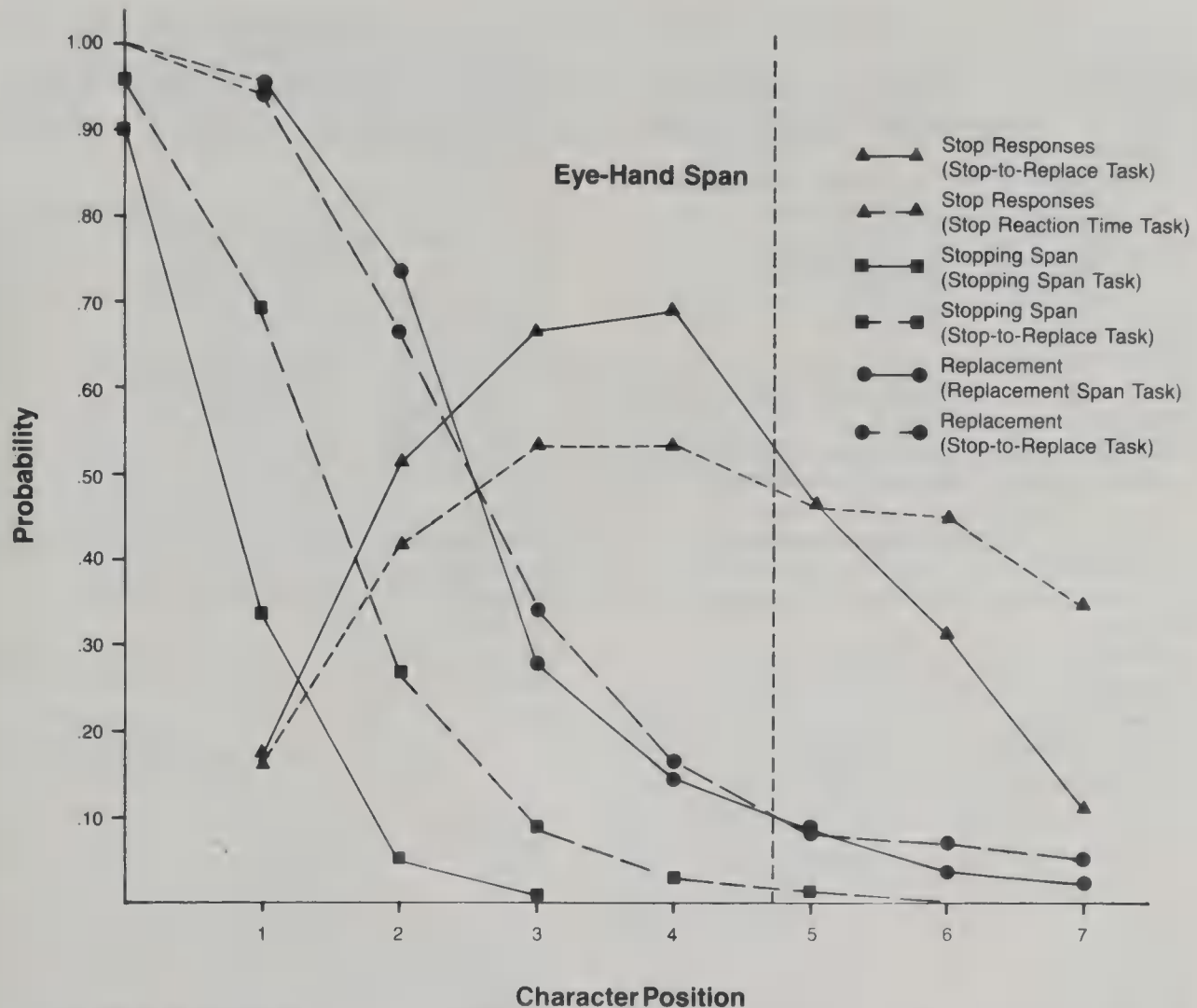


Figure 3. Probability functions across character positions for the experimental tasks in the supplemental study. (The dashed vertical line represents the mean eye-hand span across subjects.)

The present studies clearly demonstrate that multiple spans of anticipatory processing can be identified in transcription typing. Even when the same typists were examined with the same types of material, markedly different distributions of span magnitude were obtained from different procedures. There is always some risk that estimates of varying magnitude are merely artifacts of the different procedures that are necessary to assess amount of anticipation at different processing components, but this possibility is minimized in the current project by using two separate procedures to obtain estimates relevant to each hypothesized component. That is, input processing was investigated by the detection span and copy span procedures, parsing processing was assessed by the eye-hand span and replacement span procedures, and execution processing was examined with the stopping span procedure that was found to yield values comparable to those previously reported from analyses of error detection and sensitivity to constraining context.

Because the evidence for the different spans was obtained from the same typists, the spans necessarily correspond to different amounts of time during which preparatory processing is occurring. That is, the product of median interkey interval and span size in number of characters in Studies 1 and 2 averaged 1,430 ms for detection span, 1,132 ms for copy span, 868

ms for eye-hand span, 484 and 509 ms for the two estimates of the replacement span, and 224 ms for the stopping span.

These results are relevant to a hypothesis proposed by Butsch (1932) that the central nervous system appears to be organized such that processing begins approximately 1 s in advance of the required action. The present results indicate that this 1-s rule is at best very gross and at worse misleading because several distinct types of preparation can be distinguished, each apparently having its own unique temporal constants.

What do the current spans measure? Perhaps the safest conclusion is that they reflect the time course of different types of processing relevant to a specific keystroke. Processing concerned with the initial input of the material appears to occur about six to seven characters before the keystroke, whereas something analogous to the parsing or isolation of characters occurs between three to four characters in advance of the keystroke. And finally, specific keystrokes are committed about one to two characters before the actual keystroke.

References

- Butsch, R. L. C. (1932). Eye movements and the eye-hand span in type-writing. *Journal of Educational Psychology*, 23, 104-121.
- Dvorak, A., Merrick, N. L., Dealey, W. L., & Ford, G. C. (1936). *Type-writing behavior*. New York: American Book Company.

- Fendrick, P. (1937). Hierarchical skills in typewriting. *Journal of Educational Psychology*, 28, 609-620.
- Gentner, D. R. (1982). Evidence against a central control model of timing in typing. *Journal of Experimental Psychology: Human Perception and Performance*, 8, 793-810.
- Gentner, D. R. (1983). Keystroke timing in transcription typing. In W. E. Cooper (Ed.), *Cognitive aspects of skilled typewriting* (pp. 95-120). New York: Springer-Verlag.
- Logan, G. D. (1982). On the ability to inhibit complex movements: A stop-signal study of typewriting. *Journal of Experimental Psychology: Human Perception and Performance*, 8, 778-792.
- Logan, G. D. (1983). Time, information, and the various spans in typewriting. In W. E. Cooper (Ed.), *Cognitive aspects of skilled typewriting* (pp. 197-224). New York: Springer-Verlag.
- Long, J. (1976). Visual feedback and skilled keying: Differential effects of masking the printed copy and the keyboard. *Ergonomics*, 19, 93-110.
- Nelson, M. J., & Denny, E. K. (1960). *The Nelson-Denny Reading Test*. Boston: Houghton Mifflin.
- Rabbitt, P. M. A. (1978). Detection of errors by skilled typists. *Ergonomics*, 21, 945-958.
- Rothkopf, E. Z. (1980). Copying span as a measure of the information burden in written language. *Journal of Verbal Learning and Verbal Behavior*, 19, 562-572.
- Salthouse, T. A. (1984). Effects of age and skill in typing. *Journal of Experimental Psychology: General*, 113, 345-371.
- Salthouse, T. A. (1985a). Anticipatory processing in transcription typing. *Journal of Applied Psychology*, 70, 264-271.
- Salthouse, T. A. (1985b). Speed of behavior and its implications for cognition. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (2nd ed., pp. 400-426). New York: Van Nostrand Reinhold.
- Salthouse, T. A. (1986). Perceptual, cognitive, and motoric aspects of transcription typing. *Psychological Bulletin*, 99, 303-319.
- Salthouse, T. A., & Somberg, B. L. (1982). Skilled performance: Effects of adult age and experience on elementary processes. *Journal of Experimental Psychology: General*, 111, 176-207.
- Shaffer, L. H. (1975). Multiple attention in continuous verbal tasks. In P. M. A. Rabbitt & S. Dornic (Eds.), *Attention and performance, V* (pp. 435-446). New York: Academic Press.
- Shaffer, L. H., & Hardwick, J. (1969). Errors and error detection in typing. *Quarterly Journal of Experimental Psychology*, 21, 209-213.
- Somberg, B. L., & Salthouse, T. A. (1982). Divided attention abilities in young and old adults. *Journal of Experimental Psychology: Human Perception and Performance*, 8, 651-663.
- Wechsler, D. (1958). *Wechsler Adult Intelligence Scale*. Psychological Corporation: New York.
- West, L. J., & Sabban, Y. (1982). Hierarchy of stroking habits at the typewriter. *Journal of Applied Psychology*, 67, 370-376.

Received January 16, 1986

Revision received June 16, 1986 ■

Effects of Self- and Competitor Goals on Performance in an Interdependent Bargaining Task

Vandra L. Huber
Department of Management
University of Utah

Margaret A. Neale
Department of Management and Policy
University of Arizona

We investigated the relation between goal specificity and difficulty and performance on an interdependent bargaining task. In all, 102 subjects competed as buyers and sellers in a 25-min market simulation in which each negotiator was assigned either a nonspecific do-your-best objective or a specific easy, moderate, or difficult goal. Results showed that negotiators who were assigned specific, difficult goals were individually more profitable than negotiators who were assigned easier or nonspecific goals. Concerning dyadic performance, nonspecific or easy goals led to compromise agreements. Integrative agreements that benefited both parties to the transaction were facilitated by assigning both negotiators a moderate goal or difficult-moderate disparate goals. When both negotiators had difficult goals, dyadic performance did not approach the integrative level.

In recent years, one component of the bargaining and negotiation literature—the consequences of aspiration levels on the quality of bargaining outcomes—has received considerable attention (Bazerman, Magliozzi, & Neale, 1985; Huber & Neale, 1986; Neale & Northcraft, 1985). This stream of research augments much of the traditional goal-setting literature (Locke, Shaw, Saari, & Latham, 1981) by examining the influence of goals within the dyadic context of negotiation. In general, these studies consistently illustrate the generalizability and usefulness of challenging but attainable goals in achieving superior negotiated outcomes. However, with few exceptions, the goal-setting and bargaining literatures have focused on negotiation as if it were an independent task. That is, the results of previously cited research have emphasized the influence of the focal negotiator's goals and have ignored the influence of the aspiration level of the competitive other (Hamner & Harnett, 1974; Neale & Bazerman, 1985a).

Because bargaining is an interdependent process (Hamner & Harnett, 1974; Siegel & Fouraker, 1960), it is important to determine the influence of differing aspiration levels of participants as they interact in the bargaining arena. The purpose of this study was to compare the effects of different types of goals (i.e., nonspecific and specific—easy, moderate, and difficult) on individual and joint performance in an interdependent bargaining task.

Studying the effects of self- and other externally set aspiration levels on a bargaining task is critical for several reasons. First, there has been a resurgence of interest in the investigation of individual decision making when the task is interdependent. Although traditionally it has been assumed that interdependent tasks were either zero-sum propositions—what one person

gains, the other person loses—or additive, recent research has shown that interdependent tasks may lead to integrative solutions (Bazerman & Lewicki, 1983; Fisher & Ury, 1981; Pruitt, 1981). Solutions are integrative to the extent that they reconcile the parties' interests and yield joint benefits that are higher than those that could be created by a simple compromise (e.g., splitting the difference or dividing the profits equally) or by more competitive strategies (Bazerman et al., 1985). Second, because many researchers (Bazerman et al., 1985; Ben Yoav & Pruitt, 1982; Kimmel, Pruitt, Magenau, Konar-Goldband, & Carnevale, 1980; Nachajsky, Carnevale, Van Slyck, & Pruitt, 1982) have reported that integrative solutions are most likely to be achieved if both parties have high aspirations (goals), it is of particular interest to identify the conditions under which goal setting facilitates an integrative orientation.

Goal Specificity

A wealth of research on independent tasks indicates that the setting of specific, difficult goals leads to higher performance than nonspecific, do-your-best or easier goals (Locke et al., 1981). However, it is questionable whether these findings are directly applicable to interdependent bargaining tasks. Bargaining tasks specifically involve the coordination of expectations to reach agreement (Schelling, 1960). Even when one party to the transaction has a specific goal, it is unclear what effect the specificity, or lack thereof, of a second party's goal has on joint and individual bargaining outcomes.

Generalizability also is difficult for a second reason. Bargaining researchers—in contrast to goal-setting theorists—traditionally have confused goal specificity and difficulty. For example, Pruitt (1983) indicated that negotiators are more likely to behave integratively if they adopt difficult goals (high aspirations) and one or both of them adopts a problem-solving strategy. Unfortunately, his operational definition of *difficult* goal is more consistent with the definition of *easy* goals in the goal-setting literature (Locke et al., 1981), inasmuch as virtually all of the subjects achieved the stated goal. What Pruitt termed an

We gratefully acknowledge the comments of Gregory Northcraft and two anonymous reviewers on an earlier draft of this article.

Correspondence concerning this article should be addressed to Vandra L. Huber, Department of Management, College of Business, University of Utah, Salt Lake City, Utah 84112.

easy goal is more commensurate with goal theorists' definition of a nonspecific, do-your-best goal. Thus, Pruitt (1983) actually suggested that negotiators with specific, *easy goals* perform better than those assigned nonspecific goals.

Similarly, Bazerman et al. (1985) hypothesized that more difficult goals would produce integrative agreements of higher joint benefit than would less difficult goals in a competitive market. In operationally defining less difficult goals, these researchers had a do-your-best goal condition as the less difficult goal condition and a moderately difficult (68% probability of success) goal as the difficult goal. They found that negotiators who were assigned the moderate goal performed better than did those assigned a nonspecific (do-your-best) goal. Neale and Bazerman (1985a) examined four levels of goal difficulty—do-your-best, easy, moderate, and difficult. Although they found that individual negotiator performance improved with goal difficulty, they failed to address the issue of goal specificity.

In these studies the effects of specific versus do-your-best goals on joint profit and integrative behavior were completely ignored. As a result, it is unclear whether bargainers who are both assigned specific goals produce agreements of higher joint benefit than do negotiators assigned less specific or disparate (nonspecific or specific) goals. It might be speculated that when the goals of both negotiators are nonspecific, joint performance would be minimized because the energizing and directing effects of goals are missing for both negotiators. When one or both negotiators has a specific goal, performance of the dyad will be higher.

Goal Difficulty

Research on goal setting also suggests that there is a linear relation between the difficulty level of specific goals and task performance (Locke et al., 1981). Once again, however, studies on this topic have centered on independent tasks. Regardless, there is some support that the positive effects of goal difficulty on independent tasks generalizes to interdependent bargaining tasks. In an early study, Hamner and Harnett (1974) found that when negotiating the quantity and price of a fictitious commodity, buyers who chose harder goals performed better than buyers who chose easier goals. Neale and Bazerman (1985a) also suggested a linear relation between goal difficulty and the average negotiator profit per transaction.

Although these researchers suggested that specific, difficult goals facilitate performance on interdependent tasks, their studies are far from conclusive. In the Hamner and Harnett (1974) study, no information regarding goal choice is reported. Consequently, it is unclear just how difficult or how easy the self-set goals were. In addition, because only buyer data were analyzed, the effect of competitor goals on individual and dyadic performance (joint profit) is unknown. Although Neale and Bazerman (1985a) did examine the profits of both buyers and sellers, they—as Hamner and Harnett (1974)—did not examine the joint profits of the dyad. Therefore, whether negotiators who are both assigned difficult goals behave more integratively than those assigned easy or moderate goals, is not known. The aforementioned research also provides no insights into the effects of disparate goals (i.e., moderate–difficult, easy–moderate) on joint performance.

Goal Interdependence

Pruitt and others (Bazerman et al., 1985; Ben Yoav & Pruitt, 1982; Kimmel et al., 1980; Pruitt, 1983) provide some insights into which combination of specific self- and competitor goals is most likely to lead to integrative solutions. They noted that when negotiator aspirations are not sufficiently high, negotiators are too quick to make concessions and will settle for simple solutions. In contrast, if both negotiators have difficult goals, negotiators may focus on individual gain rather than on joint profitability. Under these conditions, a level of conflict would exist that may preclude the problem solving necessary for an integrative solution and high joint profit. In other instances the attainment of both negotiator difficult goals may not be possible given the constraints of the negotiation.

Thus, integrative strategies and high joint profit are most likely to occur if both negotiators have the opportunity to utilize problem-solving strategies. If both negotiators are given moderately difficult goals (as opposed to nonspecific, easy or difficult goals), then there is room for integrative behavior. Under these conditions, goals are sufficiently specific and difficult to energize performance.

Research on creativity and innovation is supportive of the view (Femina & Sopkin, 1970; Reitz, 1981) that the establishment of “realistic” goals and deadlines facilitates creativity. There also is evidence that creativity emerges when individuals are pressed to discard obvious solutions and push forward to better solutions (Maier, 1960). In the bargaining arena, this is most likely to occur when negotiators have disparate, challenging goals and objectives.

To our knowledge, only one study (Neale & Northcraft, 1985) has been conducted that explored the effects of goal disparity. In this study, either one or both negotiators were assigned a moderate goal. When goals were disparate, one negotiator had a do-your-best goal and the other a specific, moderate goal. The results showed that specific goals resulted in more profitable transactions than did do-your-best goals. It is surprising that profitability was highest when one rather than both negotiators had a specific, moderate goal. In this situation, the outcome approached the maximum integrative level. These results suggest that goal disparity rather than goal consistency will produce higher joint performance on an interdependent task.

Following this line of reasoning, it may be the case that joint profit will be higher when negotiators have disparate rather than homogeneous goals. Because the energizing effects of goals are directly related to their difficulty, this relation is likely to be moderated by the specificity and difficulty of each negotiator's goal. Thus, disparate goals will be most energizing when they are specific and when they meet but do not exceed the maximum joint outcome possible. This is likely to be the case when one negotiator has a difficult goal and the other a moderate goal. Under these conditions, both parties will have high aspirations, yet there still is sufficient potential for creative problem solving and conflict resolution. By comparison, when both negotiators have difficult goals they may drive so hard toward the attainment of their individual difficult goal that they ignore the potential for an integrative agreement. This is likely if the combined difficulty of the two negotiator's goals exceeds the maximum integrative potential of the transaction.

Relative Profits of Dyad Members

Concerning the relative profits of negotiators who are assigned consistent or disparate goals, no definitive conclusions can be made. To our knowledge, researchers have not examined the issue of relative profits of negotiators who are assigned consistent or disparate goals. However, research on independent tasks consistently has shown a linear relation between goal difficulty and performance (Locke et al., 1981). Assuming that these results generalize, it would seem that the negotiator with the relatively more difficult goal should be more profitable. Under these conditions, the negotiator with the more difficult goal would be less willing to settle for a lower individual outcome than a negotiator assigned an easier goal.

Based on the preceding discussion, a number of hypotheses have been suggested. Regarding individual negotiator performance, prior research suggests the following.

Individual Negotiator Performance

1. Individual negotiator performance will be higher when negotiators have specific, rather than nonspecific goals.
2. There will be a direct linear relation between goal difficulty and individual negotiator performance.

Joint Profitability

Generalizing from the research on independent tasks, the following are predicted on an interdependent bargaining task:

3. The assigned goals of both negotiators will have a positive and additive effect on dyadic performance.
4. When negotiator goals are homogeneous, joint profit will be highest when both negotiators have moderate goals. Joint profit will be the lowest when both negotiators have nonspecific goals.
5. When negotiator goals are disparate, joint profit will be highest when one negotiator has a difficult goal and the other has a moderate goal.

Relative Effects of Goals

Although previous studies have not examined the relative profitability of negotiator assigned disparate goals, the following is expected:

6. When goals are disparate, the negotiator with the more difficult goal will be the more profitable.

Method

Subjects

In all, 102 undergraduate students from an organizational behavior course at the University of Arizona participated in two separate free-market simulations (i.e., runs). Both runs of the simulation had equal numbers of buyers and sellers. However, slight variations in cell sizes across the goal factor were necessary because the number of subjects in each market varied.

Procedures

A bargaining task developed by Bazerman et al. (1985) was used in this study. Subjects entering the classroom were randomly assigned to

Table 1

Buyer and Seller Profit Schedules for Positively and Negatively Framed Negotiators

Profit level	Delivery time	Discount terms	Financing terms
Seller net profit schedule			
A	000	000	000
B	200	300	500
C	400	600	1,000
D	600	900	1,500
E	800	1,200	2,000
F	1,000	1,500	2,500
G	1,200	1,800	3,000
H	1,400	2,100	3,500
I	1,600	2,400	4,000
Buyer net profit schedule			
A	4,000	2,400	1,600
B	3,500	2,100	1,400
C	3,000	1,800	1,200
D	2,500	1,500	1,000
E	2,000	1,200	800
F	1,500	900	600
G	1,000	600	400
H	500	300	200
I	000	000	000

Note. Figures are in dollars.

be either a buyer or a seller and were randomly assigned a do-your-best (nonspecific), easy, moderate, or difficult goal. Because this study focused on the effects of homogeneous and disparate goals (rather than on differences in buyer and seller profits), a 4×4 experimental design was used with buyer's goal and seller's goal as the two factors.

At the beginning of the negotiation period, each bargainer was given a set of instructions describing the exercise as a free-market simulation between buyers and sellers. Subjects were told that product quality among all manufacturers (sellers) was undifferentiable and that profits were affected by only three factors: delivery terms, discount level, and financial terms. The information packet given to buyers and sellers included the same payoff table used by Bazerman et al. (1985) and shown in Table 1. For each of the three factors, there were nine associated profit levels, A through I. Negotiators saw the profit schedule for their role only.

Although it was unlikely that the terms would be accepted, buyers achieved their highest profits (\$8,000) and sellers their lowest profits (\$00) at the A-A-A; sellers achieved their highest profit (\$8,000) and buyers their lowest profit (\$00) at the I-I-I level. A profit of \$4,000 represented the 50-50 split of maximum joint payoffs; hence, this profit level represented the equality or compromise bargaining solution. A fully integrative agreement is reached when buyers and sellers maximize delivery time and financial terms, respectively, and agree on transaction terms of A-E-I. Under this agreement, the joint profit for the transaction is \$10,400, with each negotiator earning \$5,200.

Participants were told that they could complete as many transactions as possible in the fixed market period (25 min). However, a buyer (seller) could complete only one transaction with any one seller (buyer). Because an equal number of buyers and sellers existed in each market and the simulation was perfectly symmetrical, all of the negotiators in a particular market had identical profit potential.

To instigate a transaction, buyers and sellers made contact at the front of the room and proceeded to a bargaining area to engage in actual negotiating. After reaching an agreement, the buyer and sellers jointly

completed a transaction form that identified the buyer and seller and the delivery, discount, and financial terms to which both had agreed. After jointly turning in the transaction form, the buyer and seller were free to make contact for another transaction. This cyclical procedure continued until the end of the 25-min market period.

Goal-Setting Manipulation

Goal setting was manipulated by including a "confidential memorandum" in the information packets of the subjects. The memorandum, signed by the negotiator's supervisor (i.e., sales manager or head buyer) stated that it was against company policy to accept any transaction that did not meet minimum requirements (i.e., the assigned goal). The goal levels were based on the performance levels achieved by subjects in prior studies using this task (cf. Bazerman et al., 1985). Goal conditions included (a) an easy goal of achieving an average profit of \$4,000 per transaction, (b) a moderate goal of achieving an average profit of \$4,600 per transaction, and (c) a difficult goal of achieving an average profit of \$5,400 per transaction. In the nonspecific goal condition, subjects were told only to do their best.

Measures

Joint profit (sum of buyer and seller earnings during a single transaction) served as the primary dependent variable for analyses in which dyadic performance was the focus. Secondary variables examined for dyadic effects were seller profit and buyer profit. Individual performance was assessed by the negotiator's average profit per transaction.

Analytical Procedures

Prior to testing for the hypothesized relations, means and standard deviations for individual and joint profit were calculated for the 16 dyadic experimental conditions (Seller's Goal \times Buyer's Goal). In order to determine the effects of goal specificity on individual negotiator performance, a pooled regression model was used in which a dummy variable was created to represent the market session. The advantages of this procedure is that it controls for differences among the two markets (Cohen & Cohen, 1975). Because this study was not concerned with differences in buyer-seller profitability, a second dummy variable representing role assignment (sellers were coded as 1) was also used. The inclusion of this control variable was necessary because other studies (Bazerman et al., 1985; Neale & Bazerman, 1985b; Neale, Huber, & Northcraft, in press) have found a *framing* effect such that buyers earn more money than do sellers in a symmetrical market. Dummy variables for the three specific goal conditions were entered into the equation. The nonspecific goal condition was represented by 0s for all three dummy variables. To test Hypothesis 1, planned comparisons were conducted between the do-your-best and the three specific goal conditions. Hypothesis 2 was tested in two ways. First, planned comparisons between the three specific goal conditions were conducted. Next, the correlation between goal difficult and average profit was examined. Because the intervals between goal difficulty levels were not equal, actual goal levels—4,000 for easy, 4,600 for moderate, and 5,400 for difficult—were used for this analysis.

Multiple regression also was used to examine the effects of self- and competitor goals on dyadic performance. Although dummy variables for each level of buyers and sellers could have been used, interpreting the results would have been difficult because of the number of terms in the equation. For ease of interpretation, two variables—one for buyers' goal and one for sellers' goal—were created. Goal levels were coded 0 for nonspecific goal, 4,000 for easy, 4,600 for moderate, and 5,400 for difficult. In addition, an interaction term (Buyers' Goal \times Sellers' Goal)

was derived. Planned comparisons were used to test Hypotheses 4 and 5.

A sign test (Conover, 1980) was used to test Hypothesis 6. For this analysis, each of the 12 cells in which buyers and sellers had disparate goals were coded + if the negotiator with the harder goal was more profitable and – if the negotiator with the harder goal was less profitable. The distribution of outcomes was examined to determine if it were significantly different than that which would occur by chance.

Results

The means for joint profit, buyer's and seller's individual profits for the 16 experimental conditions, are reported in Table 2. As shown, joint profit was highest when the buyer had a difficult goal and the seller had a moderate goal. Under these conditions, the dyad approached ($M = \$10,168$) the integrative level of \$10,200. By comparison, joint profit was lowest when both negotiators had nonspecific goals. Here, joint profit was \$9,164. With homogenous goals, the dyad was most profitable when in the moderate goal condition ($M = \$10,091$). Regarding the individual profit of buyers and sellers, buyers were most profitable ($M = \$5,475$) when they had a difficult goal and the seller had a moderate goal. Similarly, sellers were most profitable ($M = \$5,139$) when they had a difficult goal and the buyer had a moderate goal.

Individual Performance

The impact of negotiator goals on average profit transaction can be shown formally by the following regression equation:

$$\begin{aligned} \text{APROFIT} = & \$4,818 - 36.4\text{RUN} - 386.63\text{ROLE} + \\ & (p < .01) \\ & 224.6\text{EGOAL} + 349.7\text{MGOAL} + 781.9\text{DGOAL} + e, \\ & (p > .10) \quad (p < .01) \quad (p < .001) \end{aligned}$$

where APROFIT is equal to the average earnings per transaction of negotiators, RUN equals the market session, ROLE is seller (coded as 1) or buyer, and EGOAL, MGOAL, and DGOAL are dummy variables representing the specific goal conditions. The do-your-best goal condition is represented by 0 for all three dummy variables. The equation was significant ($p < .001$) and accounted for 34% of the variance.

Hypothesis 1. In support of Hypothesis 1, negotiators who were assigned specific goals were more profitable ($M = \$5,046$) than negotiators who were told merely to do their best ($M = \$4,597$). A planned comparison between the nonspecific goal condition and the three specific (easy, moderate, and difficult) goal groups showed that this differences in earnings was highly significant ($t = 3.57, p = .001$).

Hypothesis 2. We predicted that there would be a direct linear relation between goal difficulty and individual negotiator performance. This proposition was partially supported. The correlation between goal difficult and average profit was .39 ($p < .001$). Planned comparisons indicated that the average profit of negotiators assigned a difficult goal was significantly higher ($M = \$5,371.52$) than easy goal ($M = \$4,819.04$) and moderate goal ($M = \$4,925.70$) subjects ($t = 3.9, p < .001$). A second planned comparison showed that negotiators assigned a difficult goal were more profitable during transactions than negotiators assigned a moderate goal ($t = 3.09, p < .003$). How-

Table 2
Summary of Buyer, Seller, and Joint Profits Under Various Combinations of Goals

Seller's goal	Buyer's goal							
	No goal		Easy		Moderate		Difficult	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
No goal								
Joint profit	9,164.71	1,204.00	9,506.52	1,077.94	9,183.78	1,146.62	9,340.00	1,096.57
Seller's profit	4,435.29	1,249.94	4,667.39	864.37	3,945.96	1,080.53	3,810.00	1,255.15
Buyer's profit	4,729.41	1,075.49	4,839.13	920.50	5,237.84	657.58	5,530.00	771.99
Easy								
Joint profit	9,312.50	1,199.39	9,564.29	995.06	9,630.76	836.71	9,958.33	793.68
Seller's profit	4,800.00	681.91	4,800.00	514.19	4,751.27	761.19	4,650.00	525.63
Buyer's profit	4,300.00	1,373.71	4,764.29	880.83	4,879.49	771.59	5,308.33	947.89
Moderate								
Joint profit	9,770.73	1,093.67	9,800.00	885.89	10,091.89	618.41	10,168.18	609.91
Seller's profit	4,826.82	848.83	4,913.73	607.31	4,810.81	491.47	4,693.18	317.25
Buyer's profit	4,943.90	1,204.17	4,886.27	1,113.02	5,281.00	526.43	5,475.00	526.43
Difficult								
Joint profit	9,718.18	980.59	9,692.00	847.99	10,073.91	632.61	9,425.00	1,163.42
Seller's profit	5,036.36	1,022.39	5,008.00	1,040.80	5,139.13	851.59	4,255.00	1,126.46
Buyer's profit	4,681.00	1,132.90	4,684.00	1,077.29	4,934.78	767.29	5,170.00	782.10

Note. Figures are in dollars.

ever, the difference between easy and moderate goals was not significant.

Together, the results for individual performance indicate that negotiators assigned specific goals are more profitable than negotiators who are assigned nonspecific goals. In addition, it appears that there is a general linear relation between the difficulty of the negotiator's goal and average profit per transaction.

Dyadic Performance

Hypothesis 3. To examine the effects of self- and competitor goals on joint profit, a pooled regression analysis was conducted. The impact of negotiator goals can be formally shown by the following equation:

$$\text{JOINT PROFIT} = \$9,206 - 316.65\text{RUN} + .106\text{BGOAL} + .039\text{SGOAL} + .0001\text{INTERACTION} + e,$$

($p < .001$) ($p < .002$)

where JOINT PROFIT is the total profit achieved by the dyad, RUN indicates the market session, SGOAL is the assigned goal of the seller, BGOAL is the assigned goal of the buyer, and INTERACTION is the interaction between the two variables. As the analysis shows, the main effects for market run and buyer's goal were significant. The main effect for seller's goal and the interaction were not significant. However, the overall model was highly significant ($p = .001$) and accounted for 9.6% of the variance. These results do not support Hypothesis 1. Although the difficulty of the buyer's goal had a direct affect on the profitability of the dyad, the difficulty of the seller's goals did not have a significant influence. In addition, the interaction between the two variables was not significant.

Hypothesis 4. We predicted that when negotiator goals are homogeneous, joint profit will be highest when both negotiators have moderate goals. This proposition was supported. The total

profit of buyers and sellers assigned moderate goals ($M = \$10,091$) was significantly higher ($t = 2.06$, $p = .02$) than that achieved when both negotiators had do-your-best conditions ($M = \$9,164$). In addition, the total profit of negotiators assigned moderate goals was significantly higher ($t = 2.90$, $p = .004$) than that achieved when both negotiators had easy ($M = \$9,564$) or difficult goals ($M = \$9,425$).

Hypothesis 5. As noted, the highest profit ($M = \$10,168$) was achieved when the buyer had a hard goal and the seller had a moderate goal. When the seller had a difficult goal and the buyer had a moderate goal, total profit averaged \$10,073. Consistent with Hypothesis 5, a planned comparison between these two conditions and all other conditions in which buyer-seller goals were disparate was significant ($t = 3.93$, $p = .001$). Subsequent analyses confirmed that the performance of the two difficult-moderate goal groups was higher than the joint profits achieved by negotiators assigned any other combinations of goals ($t = 5.76$, $p < .001$). A post hoc comparison did show that the joint profit of the two difficult-moderate goal groups were not significantly more profitable than the moderate-moderate goal group.

Relative Profits of Buyers and Sellers.

Hypothesis 6. Because a linear relation between goal difficulty and performance has generally been found, we argued that the negotiator with the harder goal would be more profitable. Twelve of the experimental conditions pitted against each other negotiators with disparate goals. In 11 of the 12 disparate goal conditions, negotiators with more difficult goals were more profitable ($p = .001$).

In sum, the findings suggest that negotiators are more likely to reach an integrative solution when one of two conditions exists: (a) Both negotiators have moderate goals, or (b) one negoti-

ator has a moderate goal and the other has a difficult goal. Joint profit is lowest when one or both negotiators have nonspecific goals. The findings also indicate that if goals vary among negotiators, the negotiator with the harder goal will profit more from the transaction. When goals are equivalent, buyers tend to be more profitable than sellers.

Discussion

The primary purpose of this study was to examine the relation between goal specificity, goal difficulty, and performance on an interdependent bargaining task. Specifically, the study examined how the performance of a negotiator was affected by how well another party performed his or her portion of the task. Unlike earlier studies that focused only on individual performance, this study examined individual as well as dyadic performance and explored the interactive effects of negotiator goals on the profitability of the dyad as well as the individual.

Consistent with the wealth of research on independent tasks (Locke et al., 1981), negotiators who were assigned specific goals were more profitable in their individual transactions than were negotiators who were assigned nonspecific goals. Regarding dyadic performance, the results of this study indicate that when both negotiators lack specific goals, performance of the dyad is significantly lower than performance achieved under any other combination of goals. Thus, it appears that on interdependent tasks—where the presence of a competing other may have motivated an individual to work hard even though they lacked a goal—subjects told to do-their-best do not do as well as they could during bargaining. As a consequence, they are less profitable individually as well as jointly.

The results of this study provide further empirical support that when negotiators were assigned disparate goals, the negotiator with the specific and relatively more difficult goal had a considerable competitive advantage. As a result, that negotiator was more profitable in the transaction. One reason this occurs is that goals have an energizing effect that is proportional to the goal's difficulty (Locke et al., 1981). Thus, the negotiator with the higher goal worked relatively harder and was unwilling to compromise at anything less than the assigned goal. By comparison, the negotiator with a lower goal could settle for less and still attain his or her objective.

Note that goal disparity (as operationally defined here) is conceptually different from the power imbalance as operationally defined by McAllister, Bazerman, and Fader (1986). Although these researchers found that imbalanced negotiator power destroyed the integrativeness of agreements, negotiators with disparate goals remain essentially equal in power (i.e., equal number of participants in each condition, identical time constraints, equivalent ability to accept or reject an offer). The primary difference among negotiators with disparate goals is in their specific level of expectation. Thus, one can easily imagine a negotiator with the more difficult goal "dragging" the dyad to greater joint profitability rather than destroying the integrative potential of the agreement by contentious behavior.

Regarding the effects of goals on dyadic profit, the results of this study provide a more complete understanding of the interdependent nature of the bargaining process. This study clearly indicates that joint profit is maximized when both participants

have specific goals. This finding clarifies the goal-difficulty/goal-specificity confusion of previous research (Bazerman et al., 1985; Neale & Bazerman, 1985a; Pruitt, 1981), clearly showing that integrative outcomes are more likely to occur when negotiator goals are specific.

When goals were specific, the difficulty of each negotiator's goal was found to play an important role in determining the profitability of the dyad. However, not all combinations of specific self- and competitor goals were equally effective in evoking high joint profit. When both negotiators had nonspecific (do-your-best) goals, performance was the lowest. When negotiators had homogeneous goals, dyadic performance was maximized when the goals were moderate. These findings extend the research of Pruitt (1981), indicating that negotiators completing multiple as well as single transactions, behave more integratively if they have moderately difficult, specific goals.

The finding that joint profit was higher when both negotiators had moderate rather than difficult goals, however, is inconsistent with goal-setting theory (Locke, 1968; Locke et al., 1981), which traditionally has found performance to be higher when the goal was difficult or even impossible (Garland, 1983). Garland (1983) found that subjects who were assigned goals beyond their immediate reach did not evidence any decrement in intrinsic motivation when compared with those who were assigned easier goals. Thus, even though the combined goals of both negotiators (\$5,400 each) exceeded the potential of the market (\$10,200), negotiators still should have been motivated to subsume as much of the potential profit in the bargaining space as possible. As the results show, this did not occur. It appears that they may have simply given up—settling in order to move on to other transactions.

One reason for the low profitability of the difficult-difficult dyad may be that negotiators assigned difficult goals focused on individual gain, rather than on dyadic profitability. As a result they tenaciously held out for high individual profitability at the exclusion of high joint profitability. When faced with the inevitability of a stalemate, they quit trying completely. This explanation seems likely for several reasons. First, negotiators were told that it was against company policy to accept any transactions that did not meet the minimum requirements (i.e., the assigned goal). Thus, there was pressure on these negotiators to attain their assigned goals. Second, negotiators assigned difficult goals generally were more successful. Their average profit per transaction was significantly higher than that achieved by negotiators assigned lower or nonspecific goals. Because, theoretically, they could compete against difficult goal subjects only 25% of the time, the odds favored holding out for the attainment of the difficult goal.

Because no measures of individual difference or process variables were included in this study, the preceding explanation is only speculative. However, the finding is interesting enough to warrant future research in which individual difference and process variables are included to pinpoint exactly what is going on when the joint goals of negotiators exceed the maximum potential of the market.

The results also indicate that high joint profit can be achieved when negotiators have disparate goals. Extending the findings of Neale and Northcraft (1985), our study showed that when one negotiator has a moderate goal and the other negotiator has

a difficult goal, joint profit is higher than that achieved under any other combination of disparate nonspecific or specific (easy, moderate, or difficult) goals. Inconsistent with the findings of Neale and Northcraft (1985), the profitability of this dyad was not significantly higher than that achieved when both negotiators had moderate goals.

Together the findings indicate that if the aspirations of negotiators are too high, too easy, or nonspecific, the likelihood of an integrative win-win solution decreases dramatically. It seems that when goals are easy, negotiators acquiesce and compromise too soon; when goals are difficult, negotiators tenaciously seek high personal profit and subsequently fail to settle at a mutually beneficial level. The study also suggests that an effective negotiator sets goals that meet or exceed those of the opposition. Such a strategy leads to high individual and joint profit. If, however, the competitor adopts a similar strategy, goal difficulty and conflict could escalate, resulting in a no-win deadlock. To prevent this, both negotiators must be willing to revise their goals downward so they are moderately difficult. Moderate goals appear difficult enough to propel the negotiators toward an integrative solution but easy enough to give negotiators the flexibility they need to reach a mutually beneficial settlement.

References

- Bazerman, M. H., & Lewicki, R. J. (Eds.). (1983). *Negotiating in organizations*. Beverly Hills, CA: Sage.
- Bazerman, M. H., Magliozzi, T., & Neale, M. A. (1985). Integrative bargaining in a competitive market. *Organizational Behavior and Human Performance*, 35, 294-313.
- Ben Yoav, O., & Pruitt, P. G. (1982, August). *Level of aspiration and expectation of future interaction in negotiation*. Paper presented at the annual conference of the American Psychological Association, Washington, DC.
- Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Conover, W. J. (1980). *Practical nonparametric statistics*. New York: Wiley.
- Femina, J. D., & Sopkin, C. (1970). *From the wonderful folks who gave you Pearl Harbor*. New York: Simon & Schuster.
- Fisher, R., & Ury, W. (1981). *Getting to yes*. Boston: Houghton Mifflin.
- Garland, H. (1983). Influence of ability, assigned goals, and normative information on personal goals and performance: A challenge to the goal attainability assumption. *Journal of Applied Psychology*, 68, 20-30.
- Hamner, W. C., & Harnett, D. L. (1974). Goal setting, performance and satisfaction in an interdependent task. *Organizational Behavior and Human Performance*, 12, 217-230.
- Huber, V. L., & Neale, M. A. (1986). Effects of cognitive heuristics and goals on negotiator performance and subsequent goal setting. *Organizational Behavior and Human Decision Processes*, 36, 342-365.
- Kimmel, M., Pruitt, D., Magenau, J., Konar-Goldband, E., & Carnevale, P. (1980). Effects of trust, aspiration, and gender on negotiation tactics. *Journal of Personality and Social Psychology*, 38, 9-23.
- Locke, E. A. (1968). Toward a theory of task motivation and incentives. *Organizational Behavior and Human Performance*, 3, 157-189.
- Locke, E. A., Shaw, K. N., Saari, L. M., & Latham, G. P. (1981). Goal setting and task performance 1969-1980. *Psychological Bulletin*, 90, 125-152.
- Maier, N. R. F. (1960). Screening solutions to upgrade quality: A new approach to problem solving under conditions of uncertainty. *Journal of Psychology*, 49, 217-231.
- McAlister, L., Bazerman, M. H., & Fader, P. (1986). Power and goal setting in channel negotiations. *Journal of Marketing Research*, 23, 228-237.
- Nachajsky, T., Carnevale, P., Van Slyck, M., & Pruitt, D. (1982, April). *Positive mood, aspirations, and negotiator behavior*. Paper presented at the meeting of the Eastern Psychological Association, Baltimore, MD.
- Neale, M. A., & Bazerman, M. H. (1985a). The effect of externally set goals on reaching integrative agreements in competitive markets. *Journal of Occupational Behavior*, 6, 12-32.
- Neale, M. A., & Bazerman, M. H. (1985b). The effects of framing and negotiator overconfidence on bargaining behaviors and outcomes. *Academy of Management Journal*, 28, 34-39.
- Neale, M. A., Huber, V. L., & Northcraft, G. (in press). The framing of negotiation: Context versus task frame. *Organizational Behavior and Human Decision Processes*.
- Neale, M. A., & Northcraft, G. (1985). *Decision biases, negotiation, and experts: The effect of framing and limit setting on integrative solutions with professional negotiators*. Working paper, University of Arizona.
- Pruitt, D. G. (1981). *Negotiation behavior*. New York: Academic Press.
- Pruitt, D. G. (1983). Integrative agreements: Nature and antecedents. In M. H. Bazerman & R. J. Lewicki (Eds.), *Negotiation inside organizations* (pp. 35-50). Beverly Hills, CA: Sage.
- Reitz, J. (1981). *Behavior in organizations*. Homewood, IL: Richard D. Irwin.
- Schelling, T. (1960). *The strategy of conflict*. Cambridge, MA: Harvard University Press.
- Siegel, S., & Fouraker, L. H. (1960). *Bargaining and group decision making*. New York: McGraw-Hill.

Received January 2, 1986

Revision received May 2, 1986 ■

Goal Importance, Self-Focus, and the Goal-Setting Process

John R. Hollenbeck and Charles R. Williams
Graduate School of Business Administration, Michigan State University

In this study we examine the role played by perceived goal importance and self-focus in the goal-setting process. More specifically, this study tests the interactive hypotheses that (a) task performance is a function of goal level, self-focus, and perceived goal importance; (b) goal level is a function of perceptions of past performance, self-focus, and perceived goal importance; and (c) perceptions of past performance are a function of actual past performance, self-focus, and perceived goal importance. Hierarchical regression analysis, using a sample of 88 retail salespersons, revealed empirical support for the first two hypotheses. Specifically, the variables described by control theory account for an increment of 6% and 8% of the variance explained in task performance and self-set goal level, respectively. Finally, implications for theory, practice, and future research are discussed.

Researchers have shown considerable interest recently in integrating traditional approaches to goal setting with more comprehensive self-regulation theories. One line of research used Bandura's (1982) self-regulation model (Locke, Frederick, Buckner, & Bobko, 1984), whereas another used Powers's (1973) control theoretic model (Campion & Lord, 1982; Fisher, 1983; Lord, Kernan, & Hanges, 1983; Taylor, 1983; Taylor, Fisher, & Ilgen, 1984). The differences between these two approaches to self-regulation have been laid out by Carver and Scheier (1981, pp. 124-126), who show that relative to Bandura's theory, control theory (a) places more emphasis on behavior maintenance than on behavior change, and (b) places less emphasis on the need for self- or external reinforcement to activate and maintain the self-regulation process. The purpose of the present study is to extend the line of research on goal setting and control theory by examining certain individual differences that according to the theory should influence the relations between past performance, goal level, and future performance.

The control theory model of goal setting and task performance has been described by Campion and Lord (1982). Campion and Lord set forth a dynamic goal-setting model in which both self-set goals and environmental feedback are incorporated into a performance-monitoring and performance-determining motivational system. The heart of this system is the negative feedback loop, typically illustrated with simple cybernetic systems such as thermostats.

Campion and Lord's (1982) empirical study highlighted two of the differences between traditional goal-setting approaches

and those associated with control theory. Specifically, Campion and Lord performed a longitudinal test of goal-setting processes that allowed for the examination of dynamic elements (e.g., goal change) under conditions of self-set goals with a sample of 188 students, who set goals with respect to performance on five tests during the course of one semester.

Whereas Campion and Lord's (1982) study was primarily concerned with the dynamic elements of self-set goals, the present research places emphasis on other aspects of the process suggested by control theory that Campion and Lord did not assess. One aspect of control theory that did not figure in Campion and Lord's study is the presence of multiple goals and the existence of a hierarchical system of control that specifies priorities among these goals. In Powers's (1978) empirical research, great pains were taken to ascertain the specific perceptions that were being controlled by subjects. His concept of "control quantity status" attempts to capture individual differences in how closely one set of goals is regulated relative to other sets of goals, that is, goal importance. In their study, Campion and Lord (1982) merely assumed that all subjects are closely regulating individual test performance. This omission leads to "unexpected" results, as 36% of the subjects who failed to achieve individual test goals actually raised their subsequent goals. Supplementary post hoc analyses revealed that this may have occurred because these subjects were not closely regulating individual test performance, but instead were controlling overall course performance. Campion and Lord also found that 44% of the subjects who failed to achieve their goals for the first test held their goals constant, thus suggesting that neither available control response (i.e., changing the goal to a lower level or counteracting the discrepancy by raising subsequent goals) was initiated. Apparently, these subjects were not regulating individual test performance or overall course performance very closely.

The notion of goal importance has implications for goal-setting applications. Specifically, it implies that the positive effect that specific, difficult goals have on task performance would be increased even further by attending to the degree of importance attributed to performance goals relative to other work- or even nonwork- related goals. For individuals who are closely regulating performance goals, even feedback indicating a minor devia-

Part of the data presented here were collected as part of the senior author's dissertation, and we are greatly indebted to committee members Eugene F. Stone, Kevin R. Murphy, Arthur P. Brief, Samuel Rabinowitz, and Eric Walton for their input. Dan Ilgen, Ken Wexley, Mike Moch, and John Wagner also provided valuable comments. Georgia Catchpole, Susan Gladis, Vincent Brennan, Bart Center, and Jill Krumholz, members of the host organization, also provided invaluable support during the data-collection phase of this study.

Correspondence concerning this article should be addressed to John R. Hollenbeck, Graduate School of Business Administration, Michigan State University, East Lansing, Michigan 48824.

tion between goals and performance may invoke behaviors aimed at reducing the discrepancy. For the individuals who are not closely regulating performance goals, even major deviations between goals and performance may not bring forth corrective effort. Therefore, the relation between performance goals and performance would be stronger for individuals who perceive performance goals as important, than for individuals not closely regulating such goals.

A second element of control theory not considered by Campion and Lord (1982) deals with individual differences in self-focus, a variable that has played a central theoretical role in most recent tests of control theory (Carver & Scheier, 1981). According to Carver and Scheier, an individual's attention can be directed in one of two directions: inward toward the self or outward toward the environment. When attention is directed inward, the individual is said to be engaging in self-focus, or self-attention. Stable and dispositional differences in self-focus are generally measured with a scale developed by Fenigstein, Scheier, and Buss (1975), which is discussed in greater detail in a later section.

Just as all goals are not important enough to be closely regulated, even important goals are not being regulated at all times. Carver and Scheier (1981) have stressed, for example, that the process described by the negative feedback loop operates only for individuals relatively high in self-focus. For individuals low in self-focus, the highly cognitive elements determining motivation (i.e., goals or standards, discrepancies between goals and feedback, etc.) simply will not be salient enough to influence discrepancy-reducing behavior. Performance outcomes for these individuals will reflect external environmental contingencies rather than internal personal goals. Indeed, there is substantial empirical evidence from laboratory studies documenting stronger goal-performance relations for high self-focused individuals relative to those low in self-focus (Carver, 1974, 1975; Scheier & Carver, 1980a, 1980b; Scheier, Fenigstein, & Buss, 1974). The increased salience of previous goal-feedback discrepancy episodes is also evidenced by the fact that high self-focused individuals demonstrate greater recall of past performance and goals (Cheeks, 1982; Pryor, Gibbons, Wicklund, Fazio, & Hood; Scheier, Buss, & Buss, 1978).

Although the major practical benefit of using these two individual difference variables is their potential to increase the goal difficulty-performance relation from a process perspective, they may also have implications for choice of goal level and perceptions of past performance (two key variables in most self-regulation theories). The effects of past performance on future goal or aspiration levels have been widely documented (Campbell, 1982; Cummings, Schwab, & Rosen, 1971; Frank, 1941; Fryer, 1964; Lopes, 1976; Wilstead & Hand, 1974). The possibility that the magnitude of this relation can be increased by considering the salience (i.e., self-focus) and importance of these goals, as suggested by control theory, has never been assessed in either laboratory or field settings. In addition, a major determinant of past perceptions of past performance is actual past performance, but as shown by Mabe and West (1982), the relation between these two variables is not always strong. Again, both self-focus and goal importance may be moderator variables with respect to the actual-performance-perceived-performance relation. Control theory and the laboratory research

testing control-theory predictions regarding the recall of past behavior (Cheeks, 1982; Pryor et al., 1977; Scheier et al., 1978) and proactive feedback-searching behavior (Scheier & Carver 1980a, 1980b) strongly suggest that accuracy of perceptions of past performance (i.e., the actual past performance-perceived past performance relation) will be higher for high self-focused individuals with goals perceived to be important relative to other subjects. Of the three hypotheses stated ahead, however, it should be recognized that these latter two are more speculative in nature. Although deductible from control theory, they do not share the supportive empirical base characteristic of the goal-performance hypothesis.

Summary and Hypotheses

The purpose of the present study is to provide an empirical test of the effects of goal importance and self-focus on the goal-setting process. Based on the theory and literature reviewed, three main hypotheses are advanced.

1. Future performance level is an interactive function of goal level, self-focus, and perceived goal importance, such that performance is highest where objectively difficult goals occur in conjunction with high self-focus and a high degree of perceived goal importance.
2. Goal level is an interactive function of perceptions of past performance, self-focus, and perceived goal performance, such that goal level will be highest where past performance levels are perceived to be high, self-focus is high, and the degree of perceived goal importance is high.
3. Perceptions of past performance will be more accurate (i.e., the actual past performance-perceived past performance relation will be greatest) for individuals characterized by high self-focus and a high degree of perceived goal importance.

Method

Subjects

Participants in this study were 143 salespersons employed by a major metropolitan department store located in the northeastern region of the United States. Of the participants, 31% were men. Subjects were selected so as to maximize their comparability on an objective sales volume that was standardized within selling departments. For this standardized score to be meaningful, it was deemed necessary to select only persons who within departments had equal opportunity to sell the same merchandise. For example, a shoe department structured so that one salesperson sold only Brand A shoes while another sold only Brand B was not included for study. For similar reasons, only salespersons who worked comparable time periods—full-time weekdays—were selected. All of the salespersons were paid a set salary plus a one half of 1% commission. From this subset of departments, only subjects from the larger departments (i.e., departments with more than 8 salespersons) were studied. Thus, the 143 subjects originally selected for study came from relatively large departments in which all individuals had an equal opportunity to sell identical merchandise during the weekdays. It should be clear that this selection procedure was used to maximize the construct validity of the performance criterion rather than to obtain a sample representative of some meaningful target population.

Due to subject attrition during the 7-month duration of the study, and to missing data caused by factors other than attrition, the final usable sample size for data analysis was 88. Given the interactive nature of the hypotheses, a power analysis was conducted to detect an effect

size for an interaction of .04 (i.e., an increment in explained variance of 4%) in a complete regression equation explaining 30% of the variance in the dependent variable at the .05 probability level. The power for 88 subjects was .70, slightly lower than the ideal of .80 recommended by Cohen (1969). Hence, a conservative bias (i.e., a decrease in the probability of rejecting the null hypothesis when that hypothesis is indeed false) has been introduced into the results.

An attrition analysis was performed to test whether there were any differences between subjects who provided complete data and those who did not. Five subjects with missing data provided grossly deviant figures (i.e., $+5.0$ *SD* from the mean), suggesting that they erroneously used weekly rather than daily sales figures. With these outliers removed, there were no differences on any measured variables between those who provided complete data and those who did not.

Procedure

There were three methods of data collection: through archival records, a questionnaire, and a policy-capturing exercise. Each of these three methods and the variables tapped by each are described here.

Archival Records

Actual past performance level. Actual past performance level for individuals was measured by accessing archival sources and recording total sales volume for the 3 months immediately preceding questionnaire administration. The three total monthly sales volume figures were then standardized within selling departments to remove level or dispersion differences across departments. The average of these three standardized scores became the final measure of actual past performance used for testing hypotheses. The average intertime period (i.e., stability) correlation among these three indices of performance was .63 ($p < .01$), and the internal consistency, that is, the standardized item alpha estimate of reliability for this measure, was .83.

Future performance level. The same archival sources were used to measure future performance levels. Total sales volume was recorded for the 3 months immediately after questionnaire administration. The three total monthly sales volume figures were again standardized within departments, and the mean of the three standardized scores became the measure of the individual's future performance level used for testing hypotheses. The average intertime period correlation among these three indices of performance was .62, and alpha reliability for this measure was .83.

Questionnaire

Perceptions of past performance. Subjects were asked to estimate, to the best of their ability, their average sales volume per day for each of the 3 previous months. The subjects should have had a rough idea of this figure inasmuch as they were required (in order to calculate their commission) to record their sales volume on a daily basis. This estimate was then multiplied by the number of days the subject worked that month (available from archival records), and this value became the estimate of total sales volume for the month in question. These total monthly sales volume figures were again standardized within selling departments to remove any between-department variation in level and dispersion, and the mean for the three standardized scores became the measure of perceptions of past performance used in hypothesis testing. The average interitem correlation among these indices was .46, yielding an alpha of .72.

Goal level. Subjects were asked to set goals for daily sales volume for the 3 months subsequent to questionnaire administration. These daily goals were then multiplied by the number of available work days for those months to arrive at a monthly sales volume goal. Again, these

figures were standardized within departments, and the mean of these standardized scores became the measure of goal level used in hypothesis testing. The average interitem correlation among the three indices was .87, yielding an alpha of .95.

Self-focus. The questionnaire also included a 17-item measure of self-focus developed by Fenigstein, Scheier, and Buss (1975). This measure, called the Self-Consciousness Scale, attempts to tap dispositional differences in the degree to which individuals' primary focus of attention is the self, rather than the environment. The scale contains such items as "I reflect about myself a lot," "I'm constantly examining my own motives," "I'm generally attentive to my inner feelings," "I'm usually aware of my appearance," and "I'm not all that concerned about the way I present myself." Although a detailed review of the literature is beyond the scope of this article, note that there has been substantial research demonstrating the construct validity of self-focus (Carver & Scheier, 1978; Davis & Brock, 1975; Diener, 1979; Diener, Lusk, DeFour, & Flax, 1980; Froming, Lopyan, & Walker, 1981; Froming & Walker, 1980; Geller & Shaver, 1976; Hass, 1979; Hull, 1980; Hull, Levenson, Young, & Sher, 1983; Hull & Levy, 1979; Scheier, Fenigstein & Buss, 1974) and the Self-Consciousness Scale, in particular (Carver & Glass, 1976; Turner, Scheier, Carver, & Ickes, 1978; Fenigstein et al., 1975). Evidence on the reliability and factor structure of the measure has been provided by Fenigstein et al. (1975). Evidence of convergent validity has been provided by Turner et al. (1978), who found that the scale was significantly correlated ($r \geq .40$) with the Guilford-Zimmerman Thoughtfulness Scale and the Pavio Imagery Inventory. Carver and Glass (1976) also found that the scale correlated significantly with self-monitoring (Snyder, 1974). This evidence suggests that persons high in self-focus are generally reflective, tend to create mental images when dealing with personal problems, and are highly attentive to the image of themselves portrayed to others. In lay terms, an individual on the high end of this measure would tend to be described as self-absorbed, narcissistic, or self-obsessed.

Evidence regarding the discriminant validity is also available. Turner et al. (1978) found that the scale was not significantly ($r \leq .20$) correlated with measures of social desirability (Crowne & Marlow, 1964), self-esteem (Morse & Gergen, 1970), and emotionality (Buss & Plomin, 1975). Carver and Glass (1976) showed that the scale was uncorrelated ($r \leq .20$) with measures of IQ (Otis, 1954), need for achievement (Edwards, 1957), test anxiety (Mandler & Sarason, 1952), and activity level or impulsivity (Buss & Plomin, 1975).

Finally, there is some evidence to suggest that the scale is composed of two dimensions, and distinctions are sometimes made between private and public self-consciousness. Fenigstein et al. (1975), for example, showed that some of the items loaded on a separate factor. This factor structure has not always been replicated, however, and in numerous studies the two "subscales" correlate more highly with each other ($r \geq .50$) than does the average interitem correlation within subscales. Indeed, in the present study, the two subscales correlated .67 with one another and a principle components factor analysis failed to indicate the suggested two-factor structure. Given this evidence, the 17 items composing this scale were treated as unidimensional. The internal consistency estimate of reliability for this measure was .78.

Policy-Capturing Exercise

Perceptions of goal importance. Based on previous theory (Powers, 1973; Vroom, 1964) and empirical research (Quinn & Staines, 1971), six facets of work that an individual may view as important were examined in the present study: (a) base pay, (b) job security, (c) the nature of the work itself, (d) co-worker relations, (e) job performance, and (f) supervisory practices. The importance an individual ascribes to performance relative to these six salient work-related perceptions was the main concern in the present study. Empirical support substantiating the

importance of the individual's perceptions of these six primary job facets is provided by Quinn and Staines's (1971) survey research, which indicates that approximately 60%–75% of the general U.S. population rate these aspects of their work as being "very important" to them. This is appreciably higher than the corresponding percentages for such work facets as "good hours," "geographical location," "pleasant physical surroundings," and "convenience to and from work."¹

The degree of control exercised over these work facets (an index of centrality or importance) is measured in a manner conceptually similar to Powers's (1978) "tracking method." In that study, subjects sat in front of a television monitor and attempted to track randomly moving points of colored light with a cursor. To determine which points were being regulated, the correlation between cursor movements and changes in position for the lights (occurring at random) was computed, and the point with the largest cursor-change–light-change correlation (usually over .90) was considered the controlled quantity.

To simulate such a procedure in this more complex setting, a policy-capturing exercise was used (Naylor & Wherry, 1965). This process involves regressing judgments, decisions, or ratings made by subjects on the cue values representing the information available to the individual making the judgments, decisions, or ratings. In the present study the measure of controlled quantity status presented subjects with the following instructions:

In the course of one's work life, his or her work situation often changes in one way or another. In some instances the individual has little control over such changes, while in other instances the individual has complete control over such changes. In the following section we would like to know how you would react to various types of changes in your work situation, if you had control over such changes. Below are descriptions of many possible changes in one's work situation. Read the description and indicate how hard you would try to bring about such a change or how hard you would try to avoid such a change.

For example, one scenario described a situation in which

Your present job was to change such that (a) the work itself was slightly more interesting and (b) your base pay increased slightly, but (c) your sales performance decreased substantially due to a lack of equipment or product information.

Another scenario described a situation in which

Your present job was to change such that (a) your sales performance increased slightly due to the availability of better equipment or product information but (b) your job security decreased slightly and (c) your co-workers were less friendly.

Subjects responded to 32 such scenarios.

The scenarios were constructed so that (a) each of the six facets was varied in both directions (i.e., changed for the better and worse), (b) each of the six facets varied in intensity of change (i.e., slight vs. substantial changes, weighted accordingly), and (c) facets that occurred in conjunction were randomly chosen (i.e., pay was not always paired with the nature of the work itself and performance as shown in the first example). There were no substantial correlations (i.e., $r \leq .20$) among cues (e.g., increases in base pay did not tend to occur with increases in performance); thus, there was no confounding of cues.

The scale used to record responses was a 5-point scale anchored by *I would try very hard to bring about such a change (+2)*, *I would try to bring about such a change (+1)*, *I would try neither to bring about nor avoid such a change (0)*, *I would try to avoid such a change (–1)*, and *I would try very hard to avoid such a change (–2)*.

When these changes are then regressed on subjects' ratings of how much effort they would extend to bring about or avoid such changes, the beta weights for each facet (i.e., pay, performance, security, etc.) give

an indication of the degree to which the facet is being actively regulated. Thus, a near-zero beta weight would indicate that the work facet manipulated in the scenario is not being closely regulated by the subject. On the other hand, a beta weight approaching 1.0 would indicate that this facet manipulated in the scenario is being very closely regulated. Thus, the beta weight associated with the job performance facet represents perceived importance of performance goals, relative to goals associated with other aspects of work.

Several methods were used to assess the meaningfulness and reliability of this measurement method. First, two scenarios were included as manipulation checks in which all of the cues were manipulated either in a positive direction or a negative direction. In the first instance, a response of -1 or -2 would be completely illogical, whereas in the second, a response of $+1$ or $+2$ would be completely illogical. Analysis of the responses for these two items indicated that these innocuous responses were made by 3% and 0% of the subjects, respectively. This provides some gross indication that the subjects in question understood the instructions and intent of the items.

Another means of assessing the meaningfulness of this measurement procedure was to calculate the multiple correlation squared (i.e., R^2) between the cues and the ratings for each subject. This serves as a check on the internal consistency of subjects' responses in that inconsistent responses (e.g., closely controlling a perception on one set of scenarios, whereas at the same time ignoring that same perception on a similar set) with respect to many or all the cues would result in an R^2 of near .19, which is the expected value of this R^2 given the number of predictors and observations (Cohen & Cohen, 1975). The mean R^2 across the 88 regressions (1 per subject) was .64. Thus, 64% of the variance in the effort ratings can be explained by the cues. Although this is not as high as the R^2 found in many policy-capturing studies, it is well above the expected value, suggesting some degree of internal consistency reliability in the ratings.

Another test of the reliability of these ratings was indexed by the stability of these ratings over time. According to Powers (1973) "controlled quantities" and, hence, goal importance should be relatively constant over short time periods. To assess this, 11 subjects were randomly selected to perform the policy capturing exercise a second time, approximately 1 month after the original survey administration. The median correlation among these responses across the 11 subjects, an index of test–retest reliability, was .72. This suggests some degree of stability in the ratings over time.

Data Analysis

The major hypotheses in this study are tested via hierarchical regression analysis (Cohen & Cohen, 1975). This method, when used to examine interaction effects, is frequently referred to as moderated regression (Saunders, 1956; Stone & Hollenbeck, 1984; Zedeck, 1971). In the first

¹ It should be pointed out here that the choice of these particular six perceptual quantities does not mean to imply that these are the only perceptual quantities that one may be controlling. That is, the individual may also be controlling for fatigue, amount of leisure time provided by the job, and so forth, but the purpose here is to empirically examine the process described by control theory, as opposed to generating a catalog of controlled quantities. The six perceptual quantities chosen here have frequently been investigated by researchers in organizational behavior, and there is empirical evidence to suggest that several of them (i.e., pay, co-workers, supervisor, and work itself) are empirically distinct dimensions (Drasgow & Miller, 1982). What should be noted, however, is that should any particularly relevant controlled quantity be omitted, the importance of any cue that is positively correlated with this omitted variable will be an overestimate of the actual importance of the measured variable.

Table 1
Means, Standard Deviations, and Intercorrelations Among All of the Variables

Variable	M	SD	1	2	3	4	5
1. Actual past performance	0.10	0.92	—				
2. Perception of past performance	2.01	2.20	.30**	—			
3. Goal level	2.39	2.29	.39**	.27*	—		
4. Goal importance	0.34	0.19	.08	.14	.17	—	
5. Self-focus	58.67	8.60	.00	-.17	-.11	-.12	—
6. Actual future performance	-0.02	0.90	.59**	.21*	.44**	.09	-.07

* $p < .05$. ** $p < .01$.

hierarchical step, the three independent variables are simultaneously entered as predictors in a regression in which the dependent variable serves as the criterion. In the second hierarchical step, the three 2-way cross products representing the double interactions are entered. Finally, in the third hierarchical step, the single 3-way cross-product term representing the triple interaction is entered.

Results

Descriptive Statistics

The means, standard deviations, and intercorrelations among all of the measured variables are shown in Table 1. Note that although actual performance, perceived performance, and goal level ($.21 \leq r \leq .59$) are significantly correlated, these variables did not correlate with the proposed moderators of self-focus or goal importance ($-.17 \leq r .17$).

Tests of Hypotheses

Hypothesis 1. The results relevant to Hypothesis 1 are shown in Table 2. Inspection of this table reveals support for Hypothesis 1. There is a statistically significant main effect for goal level ($R^2 = .16, p < .05$), and the three-way Goal Level \times Goal Importance \times Self-Focus interaction is also statistically significant ($\Delta R^2 = .06, p < .05$). The nature of this interaction is revealed in Figure 1, in which the relation between goal level and future performance is plotted for values of ± 1.0 SD units on self-focus and goal importance. As is evident in this figure, the nature of

the interaction is in line with the hypothesis that there is an overall tendency for high goals to be associated with high performance. This relation is particularly strong for individuals high in self-focus and goal importance. When adjusted for shrinkage (Cohen & Cohen, 1975), the main and interactive effects combine to explain 17% of the variance in an objective measure of sales performance.

Hypothesis 2. The results pertinent to Hypothesis 2 are shown in Table 3. This table reveals that there was also support for Hypothesis 2. There were statistically significant main effects for both perception of past performance ($R^2 = .11, p < .05$) and goal importance ($R^2 = .04, p < .05$). Of more importance, however, were the two statistically significant Perceptions of Past Performance \times Self-Focus and Perception of Past Performance \times Self-Focus \times Goal Importance interactions.

The nature of the three-way interaction was also in line with the hypothesis (and similar to that depicted in Figure 1), in that the perceived past performance–goal level relation is strongest for individuals high in self-focus and high in perceived goal importance ($\Delta R^2 = .08, p < .05$). The two-way interaction between perceived past performance and self-focus is also evident, in that the perceptions of past-performance–goal-level relation, averaging across levels of performance control, is substantially stronger for high as opposed to low self-focus subjects. This interaction accounted for 11% of the criterion variance. All of the variables described by the model, when adjusted for shrinkage, combine to account for 31% of the variance in goal level.

Hypothesis 3. With perceptions of past performance as the criterion, only a main effect for actual past performance ($R^2 = .08, p < .05$) is found. In contrast to Hypothesis 3, there are no significant interaction effects. In general, perceptions are in line with actual performance in that perceived performance increases as actual performance increases.

Table 2
Results of Regressing Actual Future Performance on Variables

Hierarchical step	Variable	R^2	p	ΔR^2	p of Δ
1	Goal level (GL)	.16	$p < .05$.16	$p < .05$
	Goal importance (GI)	.16	$p < .05$.00	<i>ns</i>
	Self-focus (SF)	.16	$p < .05$.00	<i>ns</i>
2	GL \times PC	.16	$p < .05$.00	<i>ns</i>
	GL \times SF	.16	$p < .05$.00	<i>ns</i>
	PC \times SF	.18	$p < .05$.02	<i>ns</i>
3	GL \times PC \times SF	.24	$p < .05$.06	$p < .05$

Note. $R^2 = .24$. $\Delta R^2 = .17$.

Discussion

The purpose of the present study was to examine the role of goal importance and self-focus on the goal-setting process. Building on past theoretical (Powers, 1973) and empirical work (Campion & Lord, 1982; Carver & Scheier, 1981), we generated and tested in a sample of salespersons three hypotheses in which standardized, objective measure of sales volume served as the measure of task performance.

The first and most important hypothesis dealt with the immediate determinants of task performance. In line with the vast

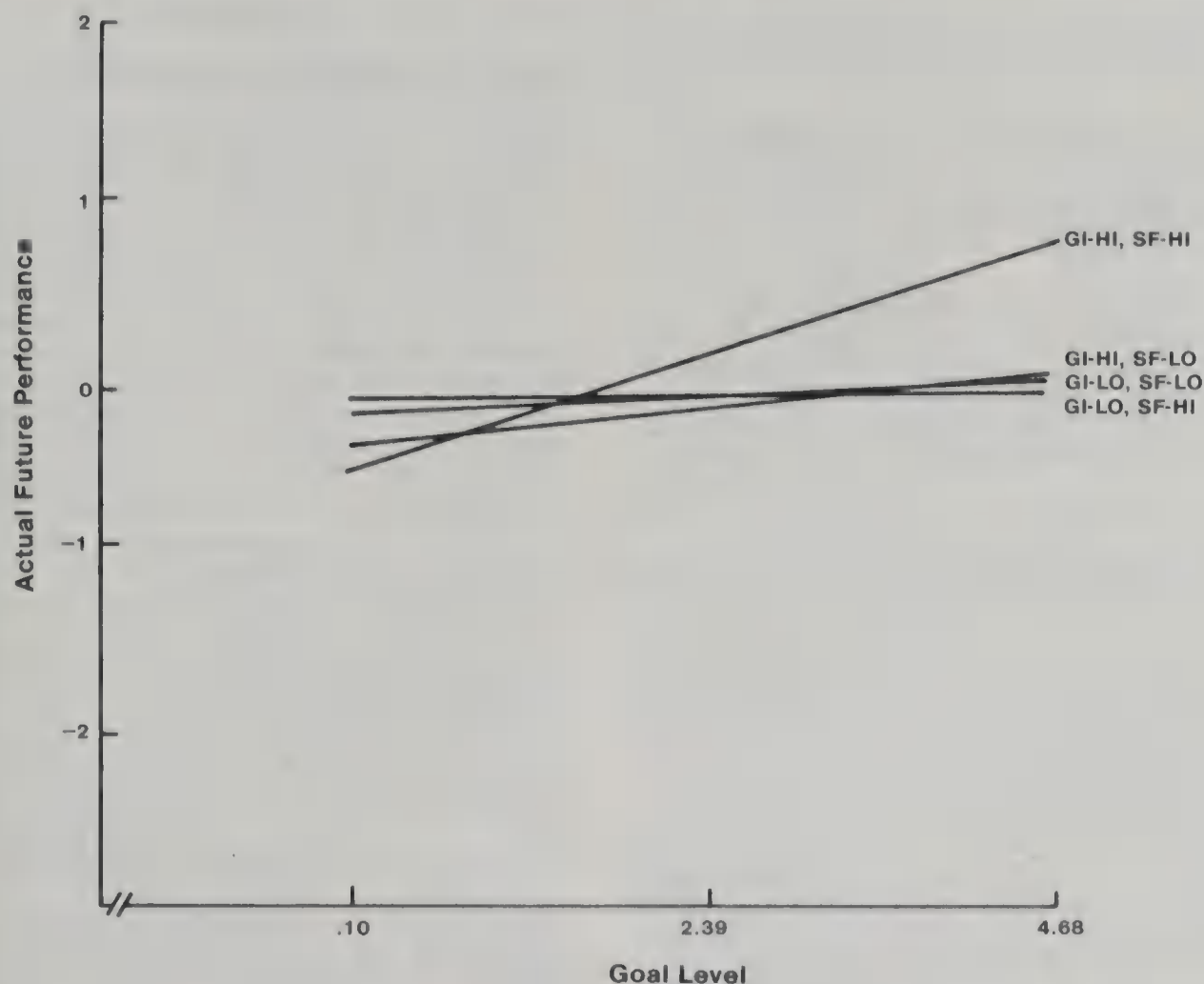


Figure 1. Triple interaction of performance control, self-focus, and goal level on future performance. (GI = goal importance; SF = self-focus; Y = predicted criterion volume; HI = high; LO = low. For GI-HI, SF-HI, $Y = .34x - .62$; for GI-HI, SF-LO, $Y = .13x - .42$; for GI-LO, SF-HI, $Y = .03x - .03$; for GI-LO, SF-LO, $Y = .05x - .08$.)

majority of empirical studies in this area (Locke et al., 1981), there was a significant goal-level main effect explaining 16% of the variance in performance. The contribution made by this study, however, is the significant interaction between goal level, self-focus, and goal importance. As predicted, the goal-level effect was significantly more pronounced for individuals characterized by high self-focus and high perceived goal importance.

The determinants of these goal levels were the primary focus of Hypothesis 2. Again, in line with a larger volume of past research, the strongest predictor of goal level was perceived past performance. The positive main effect for this variable accounted for 11% of the variance in goal level. The contribution here, however, is evidenced in the interactions of perceived past performance with self-focus and goal importance variables. For example, there was a two-way interaction between perceived past performance and self-focus ($\Delta R^2 = .11$). This interaction indicated that the effects of perceived past performance on goals was much stronger for high self-focused salespersons relative to their low self-focused counterparts. When this difference was considered in light of differences in perceived goal importance, a three-way interaction accounting for an increment of 8% of the criterion variance was also evident. In accord with control

theory predictions, the highest goal levels were set by salespersons characterized jointly as having perceptions of high past performance levels, high self-focus, and high goal importance. Thus, self-focus and goal importance would appear to be two variables that need to be added to the list of goal-level determinants previously described by Campbell (1982).

Consideration of these findings, along with consideration of the results of previous research applying control systems conceptualizations to goal setting (Campion & Lord, 1982), produces a clearer picture of the goal-setting process. Specifically, these studies suggest that the negative feedback loop (Miller, Galanter, & Pribram, 1960), the central element in control theory, may be a useful description of how the goal-setting process works. Individuals interacting with their environment operate in order to keep certain perceptions in line with referent conditions or standards associated with these perceptions. When a discrepancy between these perceptions and these standards occurs, the system is engaged and the individual operates to reduce this discrepancy. Work by Campion and Lord (1982) has highlighted the dynamic nature of this model and has emphasized that discrepancies can be eliminated not only behaviorally (e.g., by acting on the environment) but cognitively (not acting on the environment but merely changing the reference condi-

Table 3
Results of Regressing Goal Level on Variables

Hierarchical step	Variable	R ²	p	ΔR ²	p of Δ
1	Perceptions of past performance (PPP)	.11	p < .05	.11	p < .05
	Goal importance (GI)	.15	p < .05	.04	p < .05
2	Self-focus (SF)	.15	p < .05	.00	ns
	PPP × GI	.15	p < .05	.00	ns
	PPP × SF	.26	p < .05	.11	p < .05
3	GI × SF	.29	p < .05	.03	ns
	PPP × GI × SF	.37	p < .05	.08	p < .05

Note. R² = .37. ΔR² = .31.

tion). Thus, certain situations where the negative feedback loop does not appear to predict behavior (e.g., where an individual consistently performs below goal levels) can be explained by noting that these individuals are using cognitive means of discrepancy reduction. This study builds on the earlier work by highlighting two important conditions that must be considered if one is to predict behavior using the negative feedback loop. First, not all goals are closely regulated, hence the need to consider the perceived goals performance variable. Second, there are individual differences in the extent to which people engage in this process. The predictions emanating from the negative feedback loop hold especially well for individuals who are likely to be sensitized to the cognitive elements of this process, hence the need to consider the variable of self-focus. This is not to say that past performance leads to goal levels and goal levels lead to performance only where self-focus and goal importance are high. To make this second, stronger conclusion, the evidence would have to support only the triple interaction. As Tables 2 and 3 clearly show, however, there are main effects for both past performance on goal levels and goal levels on performance. Self-focus and goal importance serve mainly to augment the strength of these relations.

There are several other limitations of this study that suggest caution when interpreting these results. One major limitation deals with the external validity of these results. First, this study only dealt with one organization, in one industry, in one geographical location. Also, the method of subject selection within this organization was driven more by the desire to have a meaningful, objective measure of performance rather than to obtain a sample representative of some target population. Therefore, no claim is made that these subjects are representative of the organization as a whole. Further, the significant attrition that will inevitably result in longitudinal research at this level in the retail industry could have affected this already unrepresentative sample in unknown ways. Yet, in studies where the primary interest is in testing theory, representativeness of the sample is of relatively little concern (Berkowitz & Donnerstein, 1982; Mook, 1983). The external validity of these findings can only be established by future research in other settings, for as Cook and Campbell (1979) stated, “external validity is enhanced more by many heterogeneous small experiments, than by one or two large experiments” (p. 80).

References

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, 37, 122–148.

Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep: Some answers to the criticisms of laboratory experiments. *American Psychologist*, 37, 245–257.

Buss, A. H., & Plomin, R. A. (1975). *Temperament theory of personality development*. New York: Wiley-Interscience.

Campbell, D. J. (1982). Determinants of goal difficulty level: A review of situational and personality influences. *Journal of Occupational Psychology*, 55, 79–95.

Campion, M. A., & Lord, R. G. (1982). A control systems conceptualization of goal setting process. *Organizational Behavior and Human Performance*, 30, 265–287.

Carver, C. S. (1974). Facilitation of physical aggression through objective self-awareness. *Journal of Experimental Social Psychology*, 10, 365–370.

Carver, C. S. (1975). Physical aggression as a function of objective self-awareness and attitudes toward punishment. *Journal of Experimental Social Psychology*, 11, 510–519.

Carver, C. S., & Glass, D. G. (1976). The self-consciousness scale: A discriminant validity study. *Journal of Personality Assessment*, 40, 169–172.

Carver, C. S., & Scheier, M. F. (1978). Self-focusing effects of dispositional self-consciousness, mirror presence, and audience presence. *Journal of Personality and Social Psychology*, 36, 324–332.

Carver, C. S., & Scheier, M. F. (1981). *Attention and self-regulation: A control theory approach to human behavior*. New York: Springer-Verlag.

Cheeks, J. M. (1982). Aggression, moderator variables, and the validity of personality tests: A peer-rating study. *Journal of Personality and Social Psychology*, 43, 1254–1269.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Crowne, D. P., & Marlowe, D. (1964). *The approval motive: Studies in evaluative dependence*. New York: Wiley.

Cummings, L., Schwab, D., & Rosen, M. (1971). Performance and knowledge of results as determinants of goal setting. *Journal of Applied Psychology*, 55, 526–530.

Davis, D., & Brock, T. C. (1975). Use of first person pronouns as a function of increased objective self-awareness and prior feedback. *Journal of Experimental Social Psychology*, 11, 381–388.

Diener, E. (1979). Deindividuation, self-awareness, and disinhibition. *Journal of Personality and Social Psychology*, 37, 1160–1171.

Diener, E., Lusk, R., DeFour, D., & Flax, R. (1980). Deindividuation: Effects of group size, density, number of observers, and group member similarity on self-consciousness and disinhibited behavior. *Journal of Personality and Social Psychology*, 39, 449–459.

Drasgow, F., & Miller, H. E. (1982). Psychometric and substantive issues in scale construction and validation. *Journal of Applied Psychology*, 67, 268–290.

Edwards, A. L. (1957). *Manual for the Edwards Personal Preference Schedule* (Rev. ed.). New York: Psychological Corporation.

Fenigstein, A., Scheier, M. E., & Buss, A. H. (1975). Public and private self-consciousness: Assessment and theory. *Journal of Consulting and Clinical Psychology*, 43, 522–527.

Fisher, C. D. (1983, August). *Dysfunctional affective and behavioral reactions to negative feedback: A control theory explanation and implications for structuring organizational feedback environments*. Pa-

- per presented at the meeting of the Academy of Management, Dallas, Texas.
- Frank, J. (1941). Recent studies of the level of aspiration. *Psychological Bulletin*, 38, 218-226.
- Froming, W. J., Lopyan K., & Walker, G. R. (1981). *Self-awareness, individual differences, and adherence to public and private standards for behavior*. Unpublished manuscript.
- Froming, W. J., & Walker, G. R. (1980). *Self-awareness and public versus private standards for behavior*. Unpublished manuscript.
- Fryer, F. (1964). *An evaluation of level of aspiration as a training procedure*. Englewood Cliffs, NJ: Prentice-Hall.
- Geller, V., & Shaver, P. (1976). Cognitive consequences of self-awareness. *Journal of Experimental Social Psychology*, 12, 99-108.
- Hass, R. G. (1979, April). *A test of bidirectional focus of attention assumption*. Paper presented at the meeting of the Eastern Psychological Association, Philadelphia.
- Hull, J. G. (1980). *A self-awareness model of the causes and effects of alcohol consumption*. Unpublished manuscript.
- Hull, J. G., Levenson, R. W., Young, R. D., & Sher, K. J. (1983). Self-awareness reducing effects alcohol consumption. *Journal of Personality and Social Psychology*, 44, 461-473.
- Hull, J. G., & Levy, A. S. (1979). The organizational functions of the self: An alternative to the Duval and Wicklund model of self-awareness. *Journal of Personality and Social Psychology*, 37, 756-768.
- Locke, E. A., Frederick, E., Buckner, E., & Bobko, P. (1984). Effects of previously assigned goals on self-set goals and performance. *Journal of Applied Psychology*, 69, 694-699.
- Lopes, L. (1976). Individual strategies in goal-setting. *Organizational Behavior and Human Performance*, 15, 268-277.
- Lord, R. G., Kernen, M. C., & Hanges, P. J. (1983, August). *An application of goal theory to understanding goal commitment and intrinsic motivation*. Paper presented at the meeting of the Academy of Management, Dallas, Texas.
- Mabe, R. A., & West, S. G. (1982). Validity of self-evaluations of ability: A review and meta-analysis. *Journal of Applied Psychology*, 67, 280-288.
- Mandler, G., & Sarason, S. B. (1952). A study of anxiety and learning. *Journal of Abnormal and Social Psychology*, 47, 166-173.
- Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. New York: Holt.
- Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, 38, 1379-1387.
- Morse, S., & Gergen, K. J. (1970). Social comparison, self-consistency, and the concept of self. *Journal of Personality and Social Psychology*, 16, 264-273.
- Naylor, J., & Wherry, R. (1965). The use of simulated stimuli and the JAN technique to capture and cluster the policies of raters. *Educational and Psychological Measurement*, 28, 3-7.
- Otis, A. D. (1954). *Manual of directions for Gamma Test*. New York: Harcourt, Brace & World.
- Powers, W. T. (1973). *Behavior: The control of perception*. Chicago: Aldine.
- Powers, W. T. (1978). Quantitative analysis of purposive systems: Some spade work at the foundations of scientific psychology. *Psychological Review*, 85, 417-435.
- Pryor, J. B., Gibbons, F. X., Wicklund, R. A., Fazio, R. H., & Hood, R. (1977). Self-focused attention and self-report validity. *Journal of Personality*, 45, 514-527.
- Quinn, R. P., & Staines, G. L. (1971). *Survey of working conditions (1969-70)*. Washington, DC: U.S. Government Printing Office.
- Saunders, D. R. (1956). Moderator variables in prediction. *Educational and Psychological Measurement*, 16, 209-222.
- Scheier, M. F., Buss, A. H., & Buss, D. M. (1978). Self-consciousness, self-report of aggressiveness, and aggression. *Journal of Research in Personality*, 12, 133-140.
- Scheier, M. F., & Carver, C. S. (1980a). *Self-directed attention and the comparison of self with standards*. Unpublished manuscript.
- Scheier, M. F., & Carver, C. S. (1980b). *Effects of dispositional self-consciousness and prior outcomes on persistence at a related task*. Unpublished manuscript.
- Scheier, M. F., Fenigstein, A., & Buss, A. H. (1974). Self-awareness and physical aggression. *Journal of Experimental Social Psychology*, 10, 264-273.
- Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of Personality and Social Psychology*, 30, 526-537.
- Stone, E. F., & Hollenbeck, J. R. (1984). Some issues associated with moderated regression. *Organizational Behavior and Human Performance*, 34, 195-213.
- Taylor, M. S. (1983, August). *Control theory: A conceptual framework for integrating the performance goal and performance feedback literatures*. Paper presented at the meeting of the Academy of Management, Dallas, Texas.
- Taylor, M. S., Fisher, C. D., & Ilgen, D. R. (1984). Individuals' reactions to performance feedback in organizations: A control theory perspective. *Research in Personnel and Human Resources Management*, 2, 81-124.
- Turner, R. G., Scheier, M. F., Carver, C. S., & Ickes, W. (1978). Correlates of self-consciousness. *Journal of Personality Assessment*, 42, 285-289.
- Vroom, V. H. (1964). *Work and motivation*. New York: Wiley.
- Wilstead, W., & Hand, H. (1974). Determinants of aspiration levels in a simulated goal setting environment of the firm. *Academy of Management Journal*, 17, 172-177.
- Zedeck, S. (1971). Problems with the use of moderator variables. *Psychological Bulletin*, 76, 295-310.

Received July 1, 1986

Revision received October 24, 1986

Accepted September 29, 1986 ■

Goal Commitment and the Goal-Setting Process: Problems, Prospects, and Proposals for Future Research

John R. Hollenbeck and Howard J. Klein
Graduate School of Business Administration, Michigan State University

The purpose of this article is to examine the role of goal commitment in goal-setting research. Despite Locke's (1968) specification that commitment to goals is a necessary condition for the effectiveness of goal setting, a majority of studies in this area have ignored goal commitment. In addition, results of studies that have examined the effects of goal commitment were typically inconsistent with conceptualization of commitment as a moderator. Building on past research, we have developed a model of the goal commitment process and then used it to reinterpret past goal-setting research. We show that the widely varying sizes of the effect of goal difficulty, conditional effects of goal difficulty, and inconsistent results with variables such as participation can largely be traced to main and interactive effects of the variables specified by the model.

The major finding emanating from the widespread research on goal setting is that difficult and specific goals lead to higher levels of performance than do easy or vague goals (Locke, Shaw, Saari, & Latham, 1981). There is some evidence to suggest, however, that one or more variables may act to moderate the relation between goal difficulty and task performance. First, the *goal difficulty effect* does not always result (e.g., Motowidlo, Loehr, & Dunnette, 1978; Oldham, 1975; Organ, 1977), and when it does, the size of the effect varies widely, from .03 (Dossett, Latham, & Saari, 1980) to .77 (Locke & Bryan, 1967). Second, the effect of goal difficulty, especially in field settings, tends to be conditional on the presence or the level of other variables (Carroll & Tosi, 1970; Dossett, Latham, & Mitchell, 1979; Hall & Hall, 1976; Ivancevich & McMahon, 1977; Latham & Saari, 1979a). Third, results are highly inconsistent with respect to the role played by other key variables such as monetary incentives (e.g., Locke, Bryan, & Kendall, 1968, vs. Pritchard & Curts, 1973), participation (Dossett et al., 1979, vs. Latham & Yukl, 1976) and individual differences (French, Kay, & Meyer, 1966, vs. Searfoss & Monczka, 1973) in the goal-setting process.

Goal commitment was one of the first potential moderating variables recognized by Locke (1968), who stated that people who "stop trying when confronted by a hard task (i.e., those uncommitted to a goal) are people who have decided that the goal is impossible to reach and who no longer are trying for that goal" (p. 164). The notion that goal commitment was a necessary condition for the goal difficulty effect was central to early theorizing on goal setting, because at that time, a number of empirical studies supported achievement motivation theory

(Atkinson & Feather, 1966) and its emphasis on goals of moderate difficulty. The results of these studies, which apparently contradict goal-setting theory, were attributed by Locke (1968) to a lack of commitment to the difficult goals set in studies of achievement motivation.

Goal commitment, according to Locke et al. (1981), refers to the determination to try for a goal. Commitment implies the extension of effort, over time, toward the accomplishment of an original goal and emphasizes an unwillingness to abandon or to lower the original goal (Campion & Lord, 1982). In addition, because of the central importance of goal difficulty in determining performance, our emphasis is on commitment to difficult goals, rather than to goals in general. There is little in the literature to advocate the use of easy goals; hence, the commitment to such goals is not a major issue.

Commitment to difficult goals should also be distinguished from acceptance of difficult goals, which merely refers to the initial use of a goal assigned by another person as a referent. Goal acceptance does not necessarily imply that the individual is bound to the standard. The present review deals conceptually with goal commitment because commitment is more critical for predicting performance. For example, one can initially accept a difficult goal and yet not demonstrate subsequent commitment to that goal over time. If commitment is a necessary condition, then the effect of goal difficulty would not be forthcoming in such an instance, despite initial goal acceptance. Although these concepts are distinguishable, note that (a) there is a considerable overlap between them, (b) they have been used almost interchangeably in past research, and (c) there is not complete consensus as to the separateness of these constructs. Because they have been used interchangeably in the past, it is impossible to determine whether the original authors are using commitment or acceptance as we have defined. Therefore, for purposes of this article, any study mentioning either commitment or acceptance is treated as dealing with commitment, and hence is included here.

Given the critical role assigned to goal commitment by early researchers on goal setting, and both theoretical and empirical

The authors are indebted to Daniel Ilgen, Edwin Locke, Charles Williams, and three anonymous reviewers for helpful comments on an earlier draft of this article. The authors would also like to thank Colleen Kniffen, who provided technical assistance.

Correspondence concerning this article should be addressed to John R. Hollenbeck, Graduate School of Business Administration, Michigan State University, East Lansing, Michigan 48824.

evidence (Erez & Zidon, 1984; Locke, 1982; Organ, 1977) that suggest that goal commitment is inversely related to goal difficulty, the assessment of goal commitment should have played a prominent role in subsequent goal-setting research. Rather, in the majority of empirical studies reviewed (66 of 109, or 61%), no mention whatsoever is made of goal commitment.¹ In another 12% of these studies, goal commitment was mentioned but never empirically assessed. In the remaining studies, goal commitment has been treated in many different ways (e.g., as a main effect variable), generally inconsistent with Locke's (1968) formulation. To date only three studies have examined commitment and tested its role as a moderator of the goal-difficulty-performance relation.

The purpose of this article is to use past empirical research and expectancy theory (Vroom, 1964) to develop a model of the antecedents and consequences of commitment to difficult goals. This model is then used to reinterpret the results of past goal-setting research (where goal commitment was not explicitly considered) and to provide direction for future research. The article is organized into four subsections dealing with (a) past empirical research, (b) the expectancy theory model, (c) reinterpretation of past research, and (d) future research directions.

Past Empirical Research on Goal Commitment

Because it is our intention to develop a model of (a) the consequences of goal commitment that is consistent with Locke's (1968) moderator formulation and (b) the antecedents of commitment to difficult goals consistent with expectancy theory, only two lines of past research are examined here. First, we describe the handful of studies that have attempted to test the moderating effects of goal commitment on the goal-difficulty-performance relation, and second, we examine research that has treated goal commitment as a dependent variable.

Studies Treating Goal Commitment as a Moderator Variable

Only three studies have tested Locke's (1968) conception of goal commitment as a moderator of the goal-difficulty-task-performance relation. Erez and Zidon (1984) found a moderating effect for goal commitment on the goal-difficulty-performance relation, whereas Frost and Mahoney (1976) and Yukl and Latham (1978) did not. Explaining these inconsistent results is difficult for several methodological reasons. First, each of these studies used only a single-item measure of commitment, and therefore, differences in reliability of measurement cannot be assessed. Second, because each study used a different item, the relative validity of each becomes questionable. Third, the studies also differ in timing of measurement, in that Erez and Zidon (1984) and Yukl and Latham (1978) measured commitment prior to having subjects engage in the task, whereas Frost and Mahoney (1976) measured it after task completion. Self-reports of goal commitment collected after task completion when the subject has complete information on both the nature of the task and the outcome relative to the goal, are hardly equivalent to measures obtained prior to subjects' engaging in the task. Finally, differential range restriction could also explain the discrepant results. Erez and Zidon (1984) designed

their study to ensure variation in goal commitment, whereas Yukl and Latham (1978) stated that only 2% of subjects were not committed to their assigned goal. Frost and Mahoney (1976) failed to report the amount of variation in their measure of commitment. Suffice to say that there have been few, if any, adequate tests of Locke's original conceptualization of goal commitment as a necessary condition for the goal difficulty effect.

Studies Treating Goal Commitment as a Dependent Variable

One commonly investigated antecedent of goal commitment has been participation, with most studies hypothesizing a positive relation between these variables. A series of studies by Latham and his associates (Dossett et al., 1979; Latham & Marshall, 1982; Latham, Mitchell, & Dossett, 1978; Latham & Saari, 1979a, 1979b; Latham & Steele, 1983; Latham, Steele, & Saari, 1982) have not supported this hypothesis; it has been supported, however, in studies by Erez, Early, and Hulin (1985) and by French et al. (1966).

The studies just described have largely searched for the determinants of goal commitment without the guidance of any wide-ranging theory. Other researchers, although they did not all examine goal commitment per se, have used an expectancy theory framework (Vroom, 1964) to study the goal-setting process. Dachler and Mobley (1973), Kalb and Boyatzis (1970), Locke, Frederick, Lee, and Bobko (1984), and Steers (1975) found that the expected probability of obtaining a goal was positively related to goal commitment. Mento, Cartledge, and Locke (1980) suggested, however, that commitment may drop off at the extreme upper end of the expected probability continuum.

Valence, another major component of expectancy theory, has also been reported to be a determinant of goal commitment. All three of the relevant studies have found positive relations (Dossett et al., 1979; Mento et al., 1980; Oldham, 1975). In summary, expectancy theory may be a useful approach for increasing our understanding of the determinants of commitment to difficult goals.

Expectancy Theory Model of the Goal Commitment Process

The possibility of integrating expectancy theory and goal-setting theory via goal commitment has been recognized previously (Dachler & Mobley, 1973; Dossett et al., 1979; Kalb & Boyatzis, 1970; Mento et al., 1980; Oldham, 1975; Steers, 1975). Furthermore, Locke et al. (1981) suggested that "the factors that affect goal acceptance . . . fit easily into two major categories, which are the main components of expectancy theory" (p. 144). Locke et al. (1981) then listed variables likely to affect expectations of goal attainment and attractiveness of goal attainment. One purpose of the model presented here is to expand on this earlier work by (a) suggesting additional variables likely to be associated with either the attractiveness or expect-

¹ A table containing, describing, and categorizing the treatment of goal commitment is available from the first author, along with a complete reference list identifying the 109 studies reviewed.

tancy of goal attainment and (b) differentiating between situational and personal determinants of attractiveness and expectancy. The value of this differentiation lies in the ability to specify person by situation interactions that are the cornerstone of interactional psychology (Ekehammar, 1974; Endler & Magnusson, 1976). The recognition that such interactions may exist will be of central importance in a later section of this article. A second purpose of this model is to reinterpret results from previous goal-setting studies in which the goal difficulty effect did not emerge or was conditional on the presence of other variables. Although commitment was not directly measured in these studies, the model developed here proposes that many studies have, in fact, shown the goal commitment by goal difficulty interaction. That is, conditional results or weak effect sizes can be directly attributed to the variables measured in these studies, which this model proposes as antecedents of commitment.

Figure 1 presents a model of the antecedent factors that may enhance the commitment to difficult goals. The antecedents of commitment are broken down first by whether they affect attractiveness or expectancy, and then further delineated by whether they are of a personal or situational nature. Note that the model is meant to illustrate the way in which many variables previously studied in this area could be theoretically linked to goal commitment, and is not meant to be a comprehensive or exhaustive list. Figure 1 also suggests, in line with the formulations of Locke (1968), that the primary consequence of goal commitment is to moderate the relation between goal difficulty and task performance. Under certain conditions goal difficulty may generate a main effect. For example, when only difficult goals are established, all else being equal, greater commitment will lead to greater performance (Locke et al., 1984). However, when the entire range of goals are represented in a sample (i.e., easy, moderate, difficult) no such main effect will be in evidence.

Situational Factors Affecting the Attractiveness of Goal Attainment

Several variables discussed by Salancik (1977) in the context of organizational commitment are worth noting here. One aspect of the situation that tends to increase commitment according to Salancik is *publicness*, that is, the extent to which significant others are aware of one's goal. It is easy to abandon a goal known only to oneself. If, however, many significant others are aware of the goal, then abandoning this goal in midstream is somewhat unattractive because it makes one appear inconsistent. Empirical support for this consistency prediction can be found in nonorganizational studies by Dweck and Gilliard (1975) and Pallak and Cummings (1976).

A second important factor related to commitment mentioned by Salancik (1977) is *volition*, defined as the extent to which an individual is free to engage in a behavior. Volition should be closely associated with goal origin in that self-set goals imply volition, assigned goals imply little volition, and participatively set goals lie somewhere in between these two extremes. Assigned goals can be abandoned without appearing inconsistent inasmuch as the goal can easily be discounted as unrealistic and the person assigning it can easily be discounted as

being out of touch. Such a cognitive response is less likely when the person chooses the goal. Empirical support for this proposition comes from research by Erez et al. (1985). In both a laboratory and a field setting, they showed that groups allowed to establish their own goals exhibited higher goal commitment relative to groups that were assigned goals.

Explicitness is a third factor described by Salancik (1977) as being a key determinant of commitment. As recognized early in the goal-setting literature, vague goals are not as effective in bringing about high performance (Locke, 1968). Salancik's (1977) theory, however, provides the rationale for why these vague goals may be of little value. There are innumerable outcomes that could be consistent with a vague goal. Reducing one's weight by 1 ounce is consistent with the goal of "losing some weight this year." The number of outcomes consistent with the goal of "losing 10 pounds by August 1" is much smaller.

Many other situational factors could act to increase goal commitment by increasing the attractiveness of goal attainment. The importance of reward structures in influencing behavior has been widely documented in the field of organizational behavior. For example, the amount of authority that supervisors have in terms of rewarding or punishing subordinates may be related to goal commitment. This would be analogous to conclusions made in the feedback literature that the desire to respond to feedback is related to the power of the feedback source (Ilgen, Fisher, & Taylor, 1979). Competition may also be related to goal commitment. Pressure generated by competitive situations may increase the desire to reach a goal beyond that which would be the case in the absence of such pressures. There is even empirical evidence to suggest that competition leads to goal choices that are unrealistically high (Forward & Zander, 1971).

Personal Factors Affecting the Attractiveness of Goal Attainment

Figure 1 also specifies several personal variables that are likely to lead to greater commitment to difficult goals by increasing the attractiveness of goal attainment. These variables are classified as *personal* in the sense that variation across individuals on these dimensions stems more from factors within the individual, as opposed to factors within the situation, and deals with constructs such as needs, beliefs, attitudes, and personality traits.

Individuals with high need for achievement are characterized by McClelland (1961) as having a preference for challenging tasks, for immediate feedback, and for situations in which it is possible to take personal responsibility. Given this description, it might be predicted that individuals with high need for achievement would exhibit greater commitment to challenging goals than would those with low need for achievement.

Moreover, empirical evidence suggests that variables that are positively related to the level of goal choice (i.e., degree of difficulty) under self-set goal conditions are also related to commitment to assigned goals (Early, 1985; Hollenbeck & Brief, in press). Therefore, studies showing personal variables affecting goal-level choices will be explicitly interpreted here as providing indirect evidence that those same variables would be related to goal commitment (Campbell, 1982). Hence, studies that indicate that subjects with high need for achievement set more

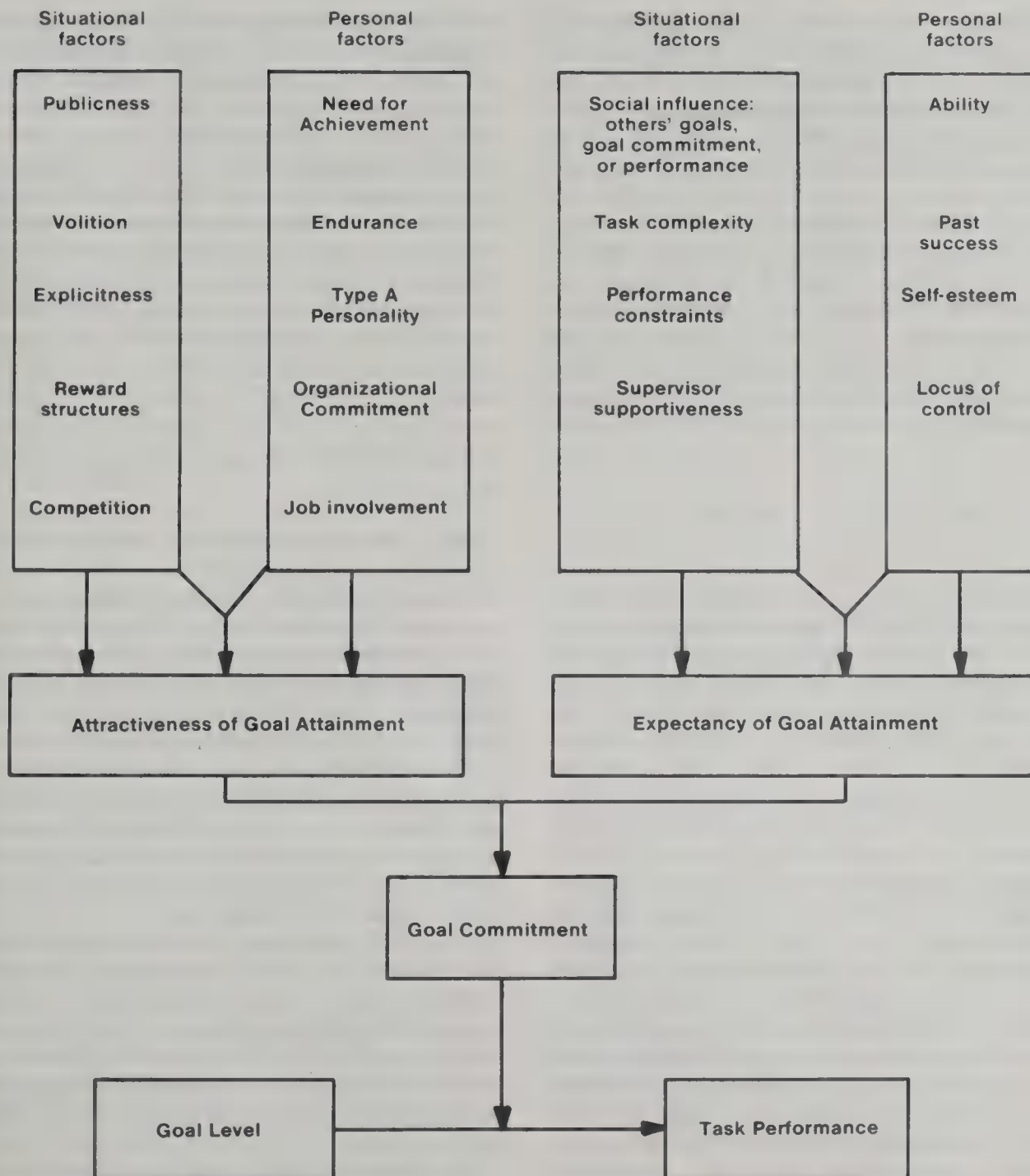


Figure 1. Expectancy theory model of the antecedents and consequences of goal commitment.

difficult goals than their counterparts with low need for achievement (Yukl & Latham, 1978) provide indirect evidence that need for achievement may be related to commitment to difficult goals.

Another personality variable that may be relevant to goal commitment is endurance. According to Jackson (1974) the individual high in endurance is "willing to work long hours; doesn't give up quickly on a problem; persevering even in the face of great difficulty; patient and unrelenting" (p. 6). Individuals high on this trait would seem much less likely to abandon difficult goals than those low in endurance.

A final personality variable that might relate to goal commitment is the degree of Type A behavior pattern. Type A people are characterized as being aggressive and competitive, setting

high standards, and putting themselves under constant time pressures (Friedman & Rosenman, 1974). It would appear that Type A individuals, when faced with difficult goals, are much more likely to redouble their efforts than to lower their goals or aspirations. Indirect evidence for this link can be found in a study by Taylor, Locke, Lee, and Gist (1984), who found that Type A individuals set more difficult goals than Type B individuals.

An individual's work-related attitudes could also be expected to relate to goal commitment by increasing the attractiveness of goal attainment. For example, organizational commitment has been defined as the relative strength of an individual's identification with and involvement in a particular organization. Given their strong identification with organizational goals in

general, it is quite likely that individuals high in organizational commitment would also be committed to operative goals established for them by the organization. Furthermore, the willingness to extend considerable effort for the organization suggests that in the face of obstacles, organizationally committed workers would tend to increase their efforts rather than modify their goals.

Job involvement is an attitude similar to organizational commitment, but the focus of identification is with the job itself rather than any one particular organization. Individuals who are highly job involved are distinguished by a strong association between their job performance and their self-esteem. Because performance on the job is central to their self-concept, highly job involved individuals would be more likely to demonstrate commitment to challenging work goals than those low in job involvement.

Situational Factors Affecting the Expectancy of Goal Attainment

One situational factor likely to have a strong impact on goal commitment because of its effects on expectancy, is social influence with respect to (a) others' performance (b) others' goals, and (c) others' goal commitment. For example, several studies have shown that knowledge of how others have performed influences self-set goal difficulty (Festinger, 1942; Lewin, Dembo, Festinger & Sears, 1944; Rakestraw & Weiss, 1981). Given this relation, it could be argued that information about others' performance would also be related to commitment to difficult goals, regardless of their origin. Indeed, Early and Kanfer (1985) manipulated the performance levels of a role model and found that subjects who viewed a high-performing role model demonstrated higher goal commitment than those who saw a low-performing model. Similarly, Bandura (1977) has shown that individuals will set high personal performance standards when they see others adopting such standards. An individual's commitment to difficult goals probably will be higher when those around him or her have similar goals, as opposed to when those around him or her have easier goals. Finally and most directly, the level of goal commitment shown by others may influence the individual's level of goal commitment. It is unlikely that an individual will maintain goal commitment when the majority of his or her co-workers are perceived as quickly abandoning goals.

Task characteristics may also be related to goal commitment. For example, on difficult or complex tasks, the link between effort and performance will not be as strong as on simple tasks. Research by Earley (1985) and Steers and Porter (1974) has documented this relation between task characteristics and goal commitment. One aspect of a task that may be critical for determining goal commitment may be the extent to which there are a number of external influences on task outcomes. Peters and O'Connor (1980) have emphasized the influence of situational constraints on work outcomes such as performance and satisfaction. The presence of such performance constraints is also likely to diminish the expectancy of goal attainment. Goal commitment would be low under such circumstances because failure to achieve the goal can be readily attributed to factors beyond the individual's control.

Finally, supervisor supportiveness may be related to goal commitment. Supervisor supportiveness was defined by Latham and Saari (1979b) as, among other things, friendliness and listening to employees' opinions. Difficult goals assigned by such supervisors will probably be perceived as fairer and more realistic, causing goal commitment to be higher. In support of this contention, Oldham (1975) directly showed that a measure of supervisory trust was positively correlated with goal commitment. Latham and Saari (1979b) provided indirect evidence in that supportive supervision was positively associated with selected goal level for subjects in participative conditions. In this latter study, the low alpha reliability (.53) associated with the commitment measure precluded detecting more direct evidence.

Personal Factors Affecting the Expectancy of Goal Attainment

Two personal variables likely to result in increased commitment to difficult goals are perceived ability and past success. The extensive literature on level of aspiration (Frank, 1941) documents the fact that, in general, future goals are higher following success than when following failure. Also, a number of studies have used self-set goals and found that self-set goal levels are strongly related to ability (Campion & Lord, 1982; Locke et al., 1984; Matsui, Okada, & Kakuyama, 1982). Similarly, subjects with high-perceived task-related ability (or self-efficacy) would be predicted to have higher expectations for achieving difficult goals, and thus higher commitment to those goals relative to low-ability subjects. Empirical support for this hypothesized relation is evidenced in a study by Locke et al. (1984) in a laboratory setting.

Hall's (1971) psychological success theory posits a linkage between self-esteem and goal commitment. Although not explicitly stated in Hall's model, the higher self-confidence that characterizes high-self-esteem persons is probably associated with high-perceived probabilities for attaining difficult goals. Empirical research on this model has shown that high self-esteem is associated with the choice of high goal levels (Hall & Foster, 1977).

Finally, locus of control may also be a personality variable that is related to goal commitment. Individuals with an external locus of control, because of their perceived inability to affect their environment, would tend to show low goal commitment. Individuals with an internal locus of control, on the other hand, when faced with a difficult goal would perceive it possible, but merely requiring more effort. Perhaps it was for this reason that in a study by Yukl & Latham (1978), employees with an internal locus of control were willing to set more difficult goals than were employees with an external locus of control.

Interactions Across Categories

The model shown in Figure 1 attempts to delineate, more concretely, the direct antecedents and consequences of goal commitment. Although only main effects have been discussed here, the possibility for interactions both within and across categories should be recognized.

Expectancy theorists (cf. Vroom, 1964) clearly predict an in-

teraction between attractiveness and expectancy; therefore, one might predict interactions between situational characteristics that differentially affect attractiveness and expectancy. For example, volition may be related to commitment only where there are not substantial constraints on performance. Where there are external constraints on performance, abandoning the original goal appears neither inconsistent nor irrational because these external factors make the original goal unrealistic.

Similar interactions could be specified between personal variables that differentially affect attractiveness and expectancy. Self-esteem, for example, may be related to goal commitment only when it occurs in conjunction with job involvement. High self-esteem individuals who are uninvolved with their work may not be as confident in being successful in the work role. Indeed, it is often found that task-specific self-esteem is a much better predictor of work behavior than is generalized self-esteem (Tharenou & Harker, 1982).

Person by situation interactions within attractiveness or expectancy categories are also possible. An interaction may exist between competition and Type A behavior pattern in determining the attractiveness of the goal attained, in that competition fails to motivate Type B individuals (Friedman & Rosenman, 1974). Expectancy of goal attainment might additionally be an interactive function of social influence and past performance. Rakestraw and Weiss (1981) found, for instance, that only subjects without prior task experience were influenced by a model's goal-setting behavior. When an individual has a long history of past performance, immediate social influence may not affect goal commitment. Finally, although not explicitly depicted in Figure 1, the possibility of feedback loops should be recognized. That is, personal and situational characteristics are not static, and task performance could have dynamic effects on our proposed antecedents such as perceived past performance, job involvement, and supervisor supportiveness.

Reinterpretation of Past Research

The importance of increasing our understanding of the role of goal commitment in the goal-setting process is highlighted by reinterpreting past research in light of the model shown in Figure 1. The lack of effects of goal difficulty, the presence of conditional effects of goal difficulty, and inconsistent results using similar key variables across studies may be, in many cases, attributable to variation in goal commitment. Note that the studies discussed ahead provide only indirect evidence for the model and are hardly a substitute for more direct empirical examinations of the linkages suggested. Their discussion will hopefully convey that the preponderance of indirect evidence on this issue warrants direct examination by future research.

Studies Failing to Establish Effects of Goal Difficulty

Several studies that found insignificant or negligible effects may be explained by subjects' lack of goal commitment. For example, in a study by Organ (1977) no group performance average reached the level of the moderate goal. Because subjects had prior task experience, they could have easily perceived the appreciably harder "difficult" goal as unrealistic and hence may have failed to become committed, because of low expectations.

Indeed, this would explain the inverse relation between goal difficulty and commitment found in this study.

Motowidlo et al. (1978) found that performance was highest for the low-goal condition (where expectancy of goal attainment was high) rather than in the high-goal condition (where expectations of goal attainment were low). In this study, however, subjects did not make their expectancy ratings conditional on trying their best. Because of this, the low expectancy reported for the difficult goals may simply reflect a lack of commitment to the originally established goal.

Ivancevich (1976) and Latham and Yukl (1975) found no significant effects of goal difficulty in field samples. In both of these cases, however, a lack of organizational support was offered as a possible explanation for these lack of results. As is shown in Figure 1, however, the effects of low organizational support were possibly mediated by a resultant decrease in goal commitment.

Finally, Oldham (1975) failed to find a significant effect of goal difficulty in a study where performance on a time-sheet completion task was the dependent variable. A substantial number of subjects did not, however, accept the assigned goals established in this study. The results from other studies that failed to replicate the effect of goal difficulty could possibly be attributed to a lack of goal commitment, but because no assessment of goal commitment was ever attempted, this is impossible to ascertain (e.g., Bavelas & Lee, 1978; Forward & Zander, 1971; Hall & Foster, 1977; Steers, 1975).

Studies Establishing Conditional Effects of Goal Difficulty

According to Locke et al. (1981), relative to laboratory studies, "the majority of the correlational [i.e., of the field] studies found only a conditional positive relation between goal difficulty and performance" (p. 129). It is instructive to look at these studies in light of the model presented here, because many of the variables on which the effect of goal difficulty was contingent are described here as antecedents of goal commitment. Therefore, whereas the situation appears to be complex or chaotic (i.e., a large number of moderating influences), the situation may be much simpler (i.e., one major moderating influence—goal commitment—that is influenced by a large number of other variables).

For example, Carroll and Tosi (1970) found an effect of goal difficulty only for self-assured managers. Such a result would be predicted from the model in Figure 1, inasmuch as self-esteem is positively related to goal commitment and goal commitment moderates the goal-difficulty-performance relation. In studies, both Ivancevich and McMahon (1977) and Steers (1975) found effects of goal difficulty only for subjects with strong higher order need strength. Because there is substantial overlap between need for achievement and higher order need strength, it is probably the case that individuals with higher order need strength simply exhibit greater goal commitment.

Goal commitment, according to the model developed here, is also higher when there are no situational constraints on performance and there is high supervisory support. It should not be surprising, therefore, that for a sample of elementary school children, an effect of goal difficulty was found only in schools offering high support (Hall & Hall, 1976). Finally, Dachler and

Mobley (1973) found an effect of goal difficulty only for long-tenured employees. There is a substantial amount of literature that suggests that tenure is associated with both job involvement (Rabinowitz & Hall, 1977) and organizational commitment (Mowday, Steers, & Porter, 1979) and, therefore, it is again apparent that the effects of goal commitment can at least partially explain an unexpected moderating effect.

Studies Yielding Inconsistent Results With Key Variables

Results of studies that have examined monetary incentives, participation, and individual differences show considerable uncertainty with respect to the roles these variables play in the goal-setting process. Locke et al. (1968) found that monetary incentives lead to setting goals of increased difficulty. This finding was not replicated by Chung and Vickery (1976); Latham et al. (1978); London and Oldham (1976); Pritchard and Curtis (1973); Terborg (1976); or Terborg and Miller (1978). Thus, one cannot conclude that monetary incentives lead to increased goal difficulty. Saari and Latham (1980) found that monetary incentives increased the frequency with which individuals set difficult goals on their own. Terborg (1976) and Terborg and Miller (1978) failed to replicate this finding, however, so it cannot be concluded that money leads to goal setting in situations wherein there would normally be no goals.

More plausibly, as suggested originally by Locke (1968) and in the model presented here, monetary incentives (or the reward structure in general) tend to increase goal commitment. Inconsistent results emerge because whereas goal commitment is a necessary condition for the goal difficulty effect, monetary incentives are not a necessary condition for goal commitment. That is, a situation may not contain monetary incentives yet still generate goal commitment through other means, such as publicness or social influence. When it is recognized that the key variable is goal commitment, it becomes apparent that one cannot make any simple prediction about the effects of monetary incentives. Rather, some assessment must be made of the relative contribution that monetary incentives make to goal commitment, over and above what is already available from other variables.

Even more inconsistent results have been obtained regarding participation. In some cases, participation resulted in higher goals (Latham et al., 1978; Latham & Yukl, 1975). In the latter study, participation interacted with individual differences to determine performance; that is, participatively set goals lead to higher performance only for educated workers (Latham & Yukl, 1975). A number of studies holding goal difficulty constant found no differences in commitment between participatively set versus assigned goals (Dossett et al., 1979; Latham & Mitchell, 1972; Latham & Saari, 1979a; Latham & Yukl, 1976). In other studies, participation seemed to play no significant role in the goal-setting process (Carroll & Tosi, 1970; Ivancevich, 1976).

The expectancy theory model of goal commitment provided here may be able to add some clarity to the situation. As discussed earlier, participation in the goal-setting process may increase volition, which in turn, may increase goal commitment. Clearly, self-set goals imply volition, whereas assigned goals do

not. Participative goal setting lies somewhere between these extremes, and without observing the actual joint goal-setting session, it is difficult to ascertain how much input subordinates had in establishing these participative goals. When the subordinate sees his or her input to be low, goal commitment will be low; when this input is perceived to be high, goal commitment will be higher (Erez et al., 1985). Again, it has to be noted that although increasing subordinate perceptions of input into the goal-setting process may be sufficient to bring about goal commitment, it is not a necessary condition for goal commitment.

According to Locke et al. (1981) "the only consistent thing about the studies of individual differences in goal setting is their inconsistency" (p. 156). The model provided here may prove instrumental in understanding the underlying reasons for these confusing results. Individual differences, according to the model, are personal factors that affect goal commitment through attractiveness or expectancy of goal attainment. Any personal factor that affects attractiveness may interact with another variable (either personal or situational) that affects expectancy and vice versa. Also, any personal factor that affects the attractiveness of goal attainment, may be substituted for by a situational variable that also affects attractiveness. For example, Steers (1975) found that goals lead to performance only for subjects with high need for achievement (i.e., a personal factor affecting attractiveness). Yet subjects with a low need for achievement performed well when they were allowed to participate in goal setting. It would appear that in this case, need for achievement was enough to bring about commitment in some subjects, but subjects low in this need required something else (in this case, volition) to bring about goal commitment. Similar explanations could explain interactions among personality and other variables found in the goal-setting literature (e.g., Latham & Yukl, 1975).

Conclusion and Suggestions for Future Research

In summary, this article supports several conclusions about the role goal commitment has played in the goal-setting literature. First, despite the fact that the earliest discussions of goal setting (Locke, 1968) specified commitment to goals as a necessary condition for goal setting to work, the majority of studies in this area have completely ignored goal commitment (or acceptance, or both). Furthermore, few investigators examined the effects of goal commitment in a fashion consistent with the conceptualization of commitment as a moderator. To date, there has not been a single study that has tested for the moderating effects of goal commitment on the goal-difficulty-performance relation in a methodologically appropriate fashion (i.e., that did not use single-item measures, measures of low reliability, range restriction, and inappropriate timing of measurement). Studies that treated goal commitment as a dependent variable, have at least provided suggestive evidence on the antecedents to commitment to difficult goals. Building on this past research, one can develop a model of the goal commitment process that breaks down the antecedents of commitment, first by determining whether they affect the attractiveness or expectancy of goal attainment, and second by determining whether they are of a personal or situational nature. When this model is used to reinterpret past goal-setting research, it becomes appar-

ent that the widely varying goal difficulty effect sizes, conditional goal difficulty effects, and inconsistent results with variables such as monetary incentives, participation, and individual differences can largely be traced to main and interactive effects of the variables specified by the model.

Future research in the area of goal setting obviously needs to place greater emphasis on assessing goal commitment. Given the central role of goal commitment in goal-setting theory, this variable should always be measured, even when the goal commitment by goal difficulty interaction is not being tested. If hypothesized goal characteristics do not affect performance, a likely explanation is that there was a low level of commitment to the goal. Direct measurement of commitment would allow this possibility to be tested and would reduce the plethora of other post hoc explanations tangential to goal-setting theory. In addition, future research, perhaps conceptually guided by the expectancy theory model of goal commitment presented, needs to uncover precisely what factors influence goal commitment. The knowledge that one has to assign "difficult" goals, has little value if one does not know how difficult these goals can be before job incumbents become uncommitted and hence abandon those goals.

Future research incorporating these recommendations may lead to a greater understanding of the role that goal commitment plays in the goal-setting process. If it is found that goal commitment is a necessary condition for goal setting to work, then this increased understanding of the antecedents of goal commitment will have critical implications for goal-setting applications, as well as goal-setting theory. Clearly this topic deserves more commitment from goal-setting researchers than has been evident in the past.

References

- Atkinson, J. W., & Feather, N. T. (1966). *A theory of achievement motivation*. New York: Wiley.
- Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice-Hall.
- Bavelas, J., & Lee, E. S. (1978). Effects of goal level on performance: A trade off of quantity and quality. *Canadian Journal of Psychology*, 32, 219–240.
- Campbell, D. J. (1982). Determinants of choice of goal difficulty level: A review of situational and personality influences. *Journal of Occupational Psychology*, 55, 79–95.
- Campion, M. A., & Lord, R. G. (1982). A control systems conceptualization of the goal setting and changing process. *Organizational Behavior and Human Performance*, 30, 265–287.
- Carroll, S. J., & Tosi, H. L. (1970). Goal characteristics and personality factors in a management-by-objectives program. *Administrative Science Quarterly*, 15, 295–305.
- Chung, K. H., & Vickery, W. D. (1976). Relative effectiveness and joint effects of three selected reinforcements in a repetitive task situation. *Organizational Behavior and Human Performance*, 16, 114–142.
- Dachler, H. P., & Mobley, W. H. (1973). Construct validity of an instrumentality-expectancy-task-goal model of work motivation. *Journal of Applied Psychology*, 58, 397–418.
- Dossett, D. L., Latham, G. P., & Mitchell, T. R. (1979). Effects of assigned versus participatively set goals, knowledge of results, and individual differences on employee behavior when goal difficulty is held constant. *Journal of Applied Psychology*, 64, 291–298.
- Dossett, D. L., Latham, G. P., & Saari, L. M. (1980). The impact of goal setting on survey returns. *Academy of Management Journal*, 23, 561–567.
- Dweck, C. S., & Gilliard, D. (1975). Expectancy statements as determinants of reactions to failure: Sex differences in persistence and expectancy change. *Journal of Personality and Social Psychology*, 32, 1077–1084.
- Earley, P. C. (1985). Influence of information, choice, and task complexity on goal acceptance, performance, and personal goals. *Journal of Applied Psychology*, 70, 481–491.
- Earley, P. C., & Kanfer, R. (1985). The influence of component participation and role models on goal acceptance, goal satisfaction and performance. *Organizational Behavior and Human Decision Processes*, 36, 378–390.
- Ekehammar, B. (1974). Interactionism in personality from a historical perspective. *Psychological Bulletin*, 81, 1026–1048.
- Endler, N. S., & Magnusson, D. (1976). *Interactional psychology and personality*. New York: Hemisphere.
- Erez, M., Early, P. C., & Hulin, C. L. (1985). The impact of participation on goal acceptance and performance: A two-step model. *Academy of Management Journal*, 28, 50–66.
- Erez, M., & Zidon, I. (1984). Effect of goal acceptance on the relation of goal difficulty to performance. *Journal of Applied Psychology*, 69, 78.
- Festinger, L. (1942). Wish, expectation, and group standards as factors influencing level of aspiration. *Journal of Abnormal and Social Psychology*, 37, 184–200.
- Forward, J., & Zander, A. (1971). Choice of unattainable group goals and effects on performance. *Organizational Behavior and Human Performance*, 6, 184–199.
- Frank, J. (1941). Recent studies of the level of aspiration. *Psychological Bulletin*, 38, 218–226.
- French, J. R., Kay, E., & Meyer, H. H. (1966). Participation and the appraisal system. *Human Relations*, 19, 3–19.
- Friedman, M., & Rosenman, R. (1974). *Type A behavior and your heart*. New York: Knopf.
- Frost, P. J., & Mahoney, T. A. (1976). Goal setting and the task process: An interactive influence on individual performance. *Organizational Behavior and Human Performance*, 17, 328–350.
- Hall, D. T. (1971). A theoretical model of career subidentity development in organizational settings. *Organizational Behavior and Human Performance*, 6, 52–76.
- Hall, D. T., & Foster, L. W. (1977). A psychological success cycle and goal setting: Goals, performance, and attitudes. *Academy of Management Journal*, 20, 282–290.
- Hall, D. T., & Hall, F. S. (1976). The relationship between goals, performance, success, self-image, and involvement under different organizational climates. *Journal of Vocational Behavior*, 9, 267–278.
- Hollenbeck, J. R., & Brief, A. P. (in press). The effects of individual differences and goal origin on the goal setting process. *Organizational Behavior and Human Decision Processes*.
- Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, 64, 349–371.
- Ivancevich, J. M. (1976). Effects of goal setting on performance and job satisfaction. *Journal of Applied Psychology*, 61, 605–612.
- Ivancevich, J. M., & McMahon, J. T. (1977). Black-white differences in a goal setting program. *Organizational Behavior and Human Performance*, 20, 287–300.
- Jackson, D. N. (1974). *Personality Research Form manual*. Port Huron, MI: Research Psychologist Press.
- Kalb, D. A., & Boyatzis, R. E. (1970). Goal setting and self-directed behavior change. *Human Relations*, 23, 439–457.
- Latham, G. P., & Marshall, H. A. (1982). The effects of self-set, partici-

- pactively set, and assigned goals on the performance of government employees. *Personnel Psychology*, 35, 399–404.
- Latham, G. P., Mitchell, T. R., & Dossett, D. L. (1978). Importance of participative goal setting and anticipated rewards on goal difficulty and job performance. *Journal of Applied Psychology*, 63, 163–171.
- Latham, G. P., & Saari, L. M. (1979a). The effects of holding goal difficulty constant on assigned and participatively set goals. *Academy of Management Journal*, 22, 163–168.
- Latham, G. P., & Saari, L. M. (1979b). Importance of supportive relationships in goal setting. *Journal of Applied Psychology*, 69, 151–156.
- Latham, G. P., & Steele, T. P. (1983). The motivational effects of participation vs goal setting on performance. *Academy of Management Journal*, 26, 406–417.
- Latham, G. P., Steele, T. P., & Saari, L. M. (1982). The effects of participation and goal difficulty on performance. *Personnel Psychology*, 35, 677–686.
- Latham, G. P., & Yukl, G. A. (1975). Assigned versus participative goal setting with educated and uneducated wood workers. *Journal of Applied Psychology*, 60, 299–302.
- Latham, G. P., & Yukl, G. A. (1976). Effects of assigned and participative goal setting on performance and job satisfaction. *Journal of Applied Psychology*, 61, 166–171.
- Lewin, K., Dembo, T., Festinger, L., & Sears, P. (1944). Level of aspiration. In J. McV. Hunt (Ed.), *Personality and the behavior disorders* (Vol. 1, pp. 33–38). New York: Ronald.
- Locke, E. A. (1968). Toward a theory of task motivation and incentives. *Organizational Behavior and Human Performance*, 3, 157–189.
- Locke, E. A. (1982). Relation of goal level to performance with a short work period and multiple goal levels. *Journal of Applied Psychology*, 67, 512–514.
- Locke, E. A., & Bryan, J. F. (1967). Performance goals as determinants of level of performance and boredom. *Journal of Applied Psychology*, 51, 120–130.
- Locke, E. A., Bryan, J. F., & Kendall, L. M. (1968). Goals and intentions as mediator of the effects of monetary incentives on behavior. *Journal of Applied Psychology*, 52, 104–121.
- Locke, E. A., Frederick, E., Lee, C., & Bobko, P. (1984). Effects of self-efficacy, goals, and task strategies on task performance. *Journal of Applied Psychology*, 69, 241–251.
- Locke, E. A., Shaw, K. N., Saari, L. M., & Latham, G. P. (1981). Goal setting and task performance: 1968–1980. *Psychological Bulletin*, 90, 125–152.
- London, M., & Oldham, G. R. (1976). Effects of varying goal types and incentive systems on performance and satisfaction. *Academy of Management Journal*, 19, 537–546.
- Matsui, T., Okada, A., & Kakuyama, T. (1982). Influence of achievement need on goal setting, performance, and feedback effectiveness. *Journal of Applied Psychology*, 67, 645–648.
- McClelland, D. C. (1961). *The achieving society*. Princeton, NJ: Van Nostrand.
- Mento, A. J., Cartledge, N. D., & Locke, E. A. (1980). Maryland vs. Michigan vs. Minnesota: Another look at the relationship of expectancy and goal difficulty to task performance. *Organizational Behavior and Human Performance*, 25, 419–440.
- Motowidlo, S. J., Loehr, V., & Dunnette, M. D. (1978). A laboratory study of the effects of goal specificity on the relationship between probability of success and performance. *Journal of Applied Psychology*, 63, 172–179.
- Mowday, R. T., Steers, R. M., & Porter, L. W. (1979). The measurement of organizational commitment. *Journal of Vocational Behavior*, 14, 224–247.
- Oldham, G. R. (1975). The impact of supervisory characteristics on goal acceptance. *Academy of Management Journal*, 18, 461–475.
- Organ, D. W. (1977). Intentional vs arousal effects of goal setting. *Organizational Behavior and Human Performance*, 18, 378–389.
- Pallak, M. S., & Cummings, W. (1976). Commitment and voluntary energy conservation. *Personality and Social Psychology Bulletin*, 2, 27–30.
- Peters, L. H., & O'Connor, E. J. (1980). Situational constraints and work outcomes: The influences of a frequently overlooked construct. *Academy of Management Review*, 5, 391–397.
- Pritchard, R. D., & Curtis, M. I. (1973). The influence of goal setting and financial incentives on task performance. *Organizational Behavior and Human Performance*, 10, 175–183.
- Rabinowitz, S., & Hall, D. T. (1977). Organizational research on job involvement. *Psychological Bulletin*, 84, 265–288.
- Rakestraw, T. L., & Weiss, H. M. (1981). The interaction of social influences and task experiences on goals, performance, and performance satisfaction. *Organizational Behavior and Human Performance*, 27, 326–344.
- Saari, L. N., & Latham, G. P. (1980). *Hypotheses on reinforcing properties of incentives contingent upon performance*. Unpublished manuscript, University of Washington.
- Salancik, G. (1977). Commitment and the control of organizational behavior and belief. In B. M. Staw & G. R. Salancik (Eds.), *New directions in organizational behavior* (pp. 1–54). Chicago: St. Claire Press.
- Searfoss, D. G., & Monczka, R. M. (1973). Perceived participation in the budget process and motivation to achieve the budget. *Academy of Management Journal*, 16, 541–554.
- Steers, R. M. (1975). Task-goal attributes, achievement, and supervisory performance. *Organizational Behavior and Human Performance*, 13, 292–303.
- Steers, R. M., & Porter, L. W. (1974). The role of task-goal attributes in employee performance. *Psychological Bulletin*, 81, 434–452.
- Taylor, M. S., Locke, E. A., Lee, C., & Gist, M. E. (1984). Type A behavior and faculty research productivity: What are the mechanisms? *Organizational Behavior and Human Performance*, 34, 402–418.
- Terborg, J. R. (1976). The motivational components of goal setting. *Journal of Applied Psychology*, 61, 613–621.
- Terborg, J. R., & Miller, H. E. (1978). Motivation, behavior, and performance: A closer examination of goal setting and monetary incentives. *Journal of Applied Psychology*, 63, 29–39.
- Tharenou, P., & Harker, P. (1982). Organizational correlates of employee self-esteem. *Journal of Applied Psychology*, 797–805.
- Vroom, V. H. (1964). *Work and motivation*. New York: Wiley.
- Yukl, G. A., & Latham, G. P. (1978). Interrelationships among employee participation, individual differences, goal difficulty, goal acceptance, goal instrumentality, and performance. *Personnel Psychology*, 31, 305–323.

Received April 25, 1986

Revision received September 17, 1986

Accepted December 1, 1986 ■

Nurse Turnover as Reasoned Action: Development of a Process Model

Perry H. Prestholdt, Irving M. Lane, and Robert C. Mathews
Louisiana State University

We used the theory of reasoned action to build a model of nurse turnover. Based primarily on the theory, a questionnaire was constructed and administered to 1,835 registered nurses. Six months after the questionnaires were completed, we obtained status information (remained or resigned) for those nurses who returned useable questionnaires. For status, differential intention was the only significant predictor. The significant predictors of differential intention were differential attitude, differential subjective norm, and differential moral obligation. For the combination of all predictors, $R^2 = .32$ for status, and $R^2 = .68$ for differential intention. These findings held up under replication procedures. Additional findings suggested potential modifications of the theory of reasoned action and the methodology used to validate its principles. Overall, the theory demonstrated its usefulness both from conceptual and applied perspectives.

For decades, researchers have been interested in the study of voluntary turnover. However, the four most comprehensive reviews in the area (Cotton & Tuttle, 1986; Hulin, Roznowski, & Hachiya, 1985; Mobley, Griffeth, Hand, & Meglino, 1979; Porter & Steers, 1973) concluded that the existing body of research has left much to be understood about the psychology of the turnover process.

Although it was not developed from the turnover research tradition, the theory of reasoned action (Ajzen & Fishbein, 1980; Fishbein, 1980) seems well suited as a conceptual base for turnover research. This theory possesses numerous characteristics listed by Mobley et al. (1979) as desirable for process models of turnover. These include (a) focusing on the individual as the unit of analysis, (b) recognizing the role of the individual's perception and evaluation of alternatives to the present job, and (c) considering the individual's intention as the immediate determinant of behavior.

The theory of reasoned action was designed to provide an understanding of complex decision-making processes. The theory assumes that people use available information in a reasonable and rational way to arrive at a behavioral decision such as withdrawal (Fishbein, 1980). Specifically, the process is viewed as a hierarchical sequence leading from beliefs, through attitudes and social norms, to intention, and finally, to behavior.

According to the theory, a person's behavioral intention (BI)

to perform (or not perform) a specific behavior is the immediate determinant of the behavior (B). A recent meta-analysis indicated the validity of this position for employee turnover (Steel & Ovalle, 1984). Intention, in turn, is determined by two constructs, the individual's personal affect or attitude toward performing (or not performing) the behavior (Aact), and the person's perception of the social influence or normative pressure to perform (or not perform) the behavior, referred to as subjective norm (SN). At this stage the theory can be expressed as a regression equation,

$$B \approx BI = w_1 Aact + w_2 SN,$$

where w_1 and w_2 are the relative weights estimated by standardized regression coefficients.

Both the affective and the normative components of a decision are based on sets of specific beliefs held by the individual. Specifically, Aact is a multiplicative function of the person's beliefs about the consequences of performing a behavior and the person's evaluation of these consequences. That is, algebraically,

$$Aact = \sum_{i=1}^n b_i e_i,$$

where b_i is the belief that performing the behavior will result in outcome i , e_i is the personal evaluation of outcome i , and n is the number of salient beliefs held by the person.

Similarly, SN is a function of the person's beliefs about what important individuals want the person to do, multiplied by the person's motivation to comply with each of these referents. That is,

$$SN = \sum_{r=1}^m NB_r Mc_r,$$

where NB_r is the belief that referent r thinks the person should or should not perform the behavior, Mc_r is the motivation to comply with the referent, and m is the number of relevant referents.

Finally, according to the theory, other factors, termed distal variables, may affect behavior, but they do so by influencing one

This research was supported by Louisiana Board of Regents' Research and Development Program Grant 83-LBR/063-B33.

We wish to thank the Louisiana Hospital Association, the Louisiana State Nurses Association, and all of the nurses and hospital administrators who gave so freely of their expertise and time. We also wish to thank four anonymous reviewers, Gregory Dobbins and Dirk Steiner for their helpful suggestions on an earlier draft, Patricia Wozniak for her help with the data analysis, and Brian Bienn, Tanya Clemons, and Elizabeth Erffmeyer for their help with data collection. We especially wish to thank Robert M. Guion for the care he devoted to this article, as evidenced by his insightful comments.

Correspondence concerning this article should be addressed to Irving M. Lane, Psychology Department, Louisiana State University, Baton Rouge, Louisiana 70803.

of the previously described components or the relative weights of the components.

Numerous research studies involving complex decisions, such as those conducted on early retirement decisions (Hwalek, Firestone, & Hoffman, 1982) and on abortion decisions (Smetana & Adler, 1980), have repeatedly indicated the effectiveness of the Fishbein and Ajzen (1975) theory. However, there have been few attempts to apply the theory to the area of voluntary turnover. Two exceptions are the research of Newman (1974) and the research of Hom and Hulin (1981).

Newman (1974) used an earlier version of the theory (Fishbein, 1967) to predict voluntary turnover of nursing home employees. Using only two of the theoretical components, Aact and NB, he accounted for 13% of the variance in voluntary turnover. This compared very favorably to 4% of the variance accounted for by an unweighted sum of the five Job Descriptive Index scales (JDI; Smith, Kendall, & Hulin, 1969) and 7% of the variance accounted for by the five JDI scales plus the General Motors Faces Scale (Kunin, 1955). Newman's research provided the initial support for Fishbein's approach in a field setting. However, because of recent advances in the definition and measurement of the theoretical components (Ajzen & Fishbein, 1980), it appears to be an inadequate test of the current version of the theory for the study of voluntary turnover.

A more recent and expansive application of Fishbein and Ajzen's (1975) approach to the area of turnover was conducted by Hom and Hulin (1981). Their research involved a competitive test of several approaches for predicting the decision of Army national guardsmen to reenlist. Among the approaches they investigated were Fishbein and Ajzen's (1975) model, Triandis's (1975) model, organizational commitment (Porter, Steers, Mowday, & Boulian, 1974), and job satisfaction as measured by the JDI (Smith et al, 1969). Their results indicated that both the Fishbein and Ajzen model and the Triandis model strongly predicted reenlistment (both R^2 s about .50), whereas job satisfaction and organizational commitment were moderate predictors of reenlistment.

Although Hom and Hulin's (1981) results again demonstrated the comparative effectiveness of Fishbein and Ajzen's (1975) approach, they did so in a situation that may not be directly comparable to the typical voluntary turnover decision. For example, the decision of the national guardsmen not to reenlist did not involve their permanent job or their major source of income. The decision also did not involve a change of employers, but rather represented alternative ways to spend one weekend a month. In addition, the reenlistment decision had to be made at a specific point in time, at the end of a contract period, rather than at any time, which is more typical of employee withdrawal decisions. Finally, by using only those guardsmen who had to make a reenlistment decision, Hom and Hulin obtained a turnover rate of 45%, nearly optimal for research purposes, but unlikely to occur in most organizational settings.

Therefore, although both Newman's (1974) and Hom and Hulin's (1981) research have provided initial evidence of the value of applying the theory of reasoned action to the study of voluntary turnover, the need for additional research seems clear. As a result, the present research tested Fishbein's (1980) theory in an organizational setting wherein the employee is making an occupational decision and is free to resign at any

time. It used all of the components in the most recent version of the theory and the measurement procedures currently recommended by Ajzen and Fishbein (1980) and Fishbein (1980).

The occupational group selected for the present research was registered nurses (RNs). They have been the focus of recent turnover investigations partly because of the estimated cost of over \$2,000 to replace an RN (Seybolt, Pavett, & Walker, 1978) and also because their relatively high turnover rate has been blamed for reducing the overall quantity and quality of patient care (Wolf, 1981).

Most of the original nursing turnover studies used a concurrent approach (e.g., Wandelt, Pierce, & Widdowson, 1981). As a result, Price and Mueller (1981) made a significant contribution to nurse turnover research when they used a longitudinal design to develop a causal model of RN turnover. Their model accounted for 18% of the variance in turnover, with most of the variance being accounted for by one factor, intent to stay. Their predictors, derived from traditional turnover research, accounted for 24% of the variance for intent to stay.

Other models of nursing turnover have met with varying degrees of success. For example, Hom, Griffeth, and Sellaro (1984), using Mobley's (1977) model of turnover, accounted for 17% of the variance in nurse turnover. In addition, Sheridan and Abelson (1983) applied a cusp catastrophe model to the nurse turnover area. Despite numerous supportive findings, their best-fit model accounted for 2% of the variance in observed turnover data.

Purpose of the Present Research

Because of both the persistent interest and the moderate success in predicting RN turnover, it would be appropriate to apply the theory of reasoned action to this problem. Therefore, the purposes of the present research were to test and replicate the theory and to evaluate two modifications that were intended to maximize its predictive effectiveness. The first modification was including moral obligation as a potential predictor of intention. It was included because several researchers (Pomazal & Jaccard, 1976; Triandis, 1975) have suggested that a personal normative component, such as moral obligation, should be added to the model. The relevance of moral obligation for a behavioral decision is dependent on the characteristics of the individuals and the behavior under investigation (Triandis, 1975). Given the altruistic history and philosophy of hospital-based nursing education (Kramer, 1981), it seems reasonable to assume the relevance of feelings of moral obligation for nurses' decisions.

The second modification was based on the recognition that the turnover decision involves two alternatives—the option to remain in the present job and the option to resign. Thus, differential measures (the difference between remaining and resigning) were used for all theoretical components. Fishbein (1980) argued that to accurately predict and fully understand the behavioral decision, it is often necessary to consider a person's belief, attitude, and intention with respect to all of the available alternatives. However, Fishbein (1980) also acknowledged that in the case of a decision between two mutually exclusive and exhaustive alternatives (e.g., to quit or not quit), differential measures of one alternative may be sufficient.

Although differential measures were not used in previous applications of the theory of reasoned action to turnover (e.g.,

Hom & Hulin, 1981), the few studies (e.g., Fishbein, 1980) that have used differentials in other behavioral domains suggested that there may be some advantages to using them. Whether or not differentials strengthen the relations between components of a turnover model is an issue of both theoretical and practical concern. Therefore, this study will conduct a direct comparison of differential measures and measures of a single behavioral alternative, resignation.

Method

Subjects

A sampling procedure based on hospital characteristics (location: rural and urban; control: profit, not for profit, and government; and size: small—less than 100 beds, medium—between 100 and 200 beds, and large—greater than 200 beds) was used to obtain a representative sample of 21 Louisiana hospitals. A total of 1,835 registered nurses, including staff nurses, administrators, and educators employed by these hospitals, received the research questionnaire. A total of 942 questionnaires were returned, of which 885 had complete data and were received in a useable time period (within 2 months). Of these, 450 were randomly selected for the initial analyses and 435 were assigned to a holdback sample for replication purposes. Status information could not be obtained on 12 nurses, and another 6 involuntarily left the employment of their hospital. Therefore, these 18 nurses were eliminated from further data analyses, reducing the number to 441 for the initial sample and 426 for the holdback sample. A total of 101 nurses (11.6%) resigned during the 6 months of the study.

Measures

The questionnaire distributed to the nurses was designed to measure all of the theoretical components as well as to obtain biographical information about the nurses. All of the questions concerned with components of the theory contained the time-constraining phrase “in the next 6 months” and were assessed with 7-point bipolar scales.

Behavioral intention was measured for two different behaviors, “remaining on the staff of this hospital” and “resigning from this hospital.” Both of the items asked the nurses to state their intention on a likely-unlikely scale. The difference between these two intention measures is referred to as differential intention. Attitude toward the act was obtained by having the nurses rate two items on a good-bad scale. The first item asked the nurses to rate “remaining on the staff of this hospital,” and the second asked them to rate “resigning from this hospital.” A differential attitude score was obtained by subtracting the latter rating from the former.

Subjective norm was measured by asking nurses the extent to which they felt that “people who are most important to me think I should” or “should not remain on the staff of this hospital.” A similar question using the should-should not scale was used to obtain the nurse’s perception of the social pressure to resign from the hospital. Differential subjective norm was obtained by determining the difference between these two scores.

For each of the 29 beliefs about resigning, nurses were asked to rate the likelihood that a specific consequence would result if they resigned from the hospital. A parallel set of 29 belief statements and likely-unlikely scales assessed their beliefs about the consequences of staying on the hospital staff. The nurses’ evaluation of each of the 29 consequences was assessed with a good-bad scale. To obtain the differential belief score for each specific consequence, the belief for resigning was subtracted from the belief for staying, and this difference was multiplied by the appropriate evaluation. The sum of these differential belief scores is referred to as differential behavioral beliefs. To measure normative belief, nurses indicated the extent to which they felt each of eight specific

referents (e.g., spouse, co-workers) thought they should or should not resign from the hospital. Motivation to comply (Mc) was assessed by asking the nurses to indicate “how much do you want to do” what each of these specific referents think you should do. The product of NB and Mc was obtained for each referent and summed across all referents. This sum was referred to as normative beliefs.

To measure moral obligation, nurses were asked to indicate if they felt they “have a moral obligation to remain on the staff of this hospital.” Similarly, they were asked to indicate on a yes-no scale the extent to which they felt a moral obligation to resign from the hospital. The difference between these two scores was obtained to determine differential moral obligation. Alternate job opportunities was assessed by asking the nurses to indicate “how likely is it that you could find another nursing job,” on a 7-point scale. This measure was conceptually similar to “expectancy re: attaining alternative” or the expectancy of being able to attain the alternative role (Mobley et al., 1979; Hom et al., 1984).

Procedure

Prior to the construction of the questionnaire, elicitation interviews were conducted with 109 nurses to obtain information on the salient beliefs, outcomes, and referents of nurses. Using information obtained from this pilot study, as well as information obtained through a search of the nurse turnover and psychological literature, the questionnaire was constructed. It, along with an explanatory cover letter, was personally distributed by the researchers to all of the nurses, who were requested to return the completed questionnaires by mail to the researchers.

Behavioral Criterion

Six months after the questionnaires were distributed, the hospitals supplied information on the employment status of each nurse. (Employment status was coded 1 if the nurse resigned or 0 if the nurse remained.)

Results

Tests of the Model

Table 1 presents the intercorrelation matrix for the 13 predictor variables (components of the model, biographical information, characteristics of the hospital) and employment status. It also contains the means and standard deviations for each variable. Table 1 indicates that all of the components of the model were significantly correlated with employment status. It also indicates that the magnitude of the correlations of the other predictors with employment status were more modest. However, both tenure and full- versus part-time status were significantly correlated with employment status.

The 13 predictors were entered into a hierarchical multiple regression analysis. First, employment status was regressed in a sequential fashion on the 11 predictors in the order they appear in Table 1. Then, differential intention was regressed in a sequential fashion on the remaining predictors in the same manner. Table 2 summarizes the results of the analysis for the initial sample, holdback sample, and the combination of these samples.

Regarding the initial sample, the results indicate that for the combination of all predictors, R^2 was .32 for employment status. Consistent with the theory, differential intention was the only significant predictor for employment status. When predicting differential intention, the combination of predictors explained 68% of the variance. Consistent with expectations,

Table 1
Means, Standard Deviations, and Intercorrelations of Variables

Variable	<i>M</i>	<i>SD</i>	1 ^a	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Employment status	0.13	0.34	—													
2. Intention ^b	3.43	4.02	-.82**	—												
3. Attitude ^b	2.61	3.63	-.76**	.81**	—											
4. Subjective norm ^b	2.76	3.53	-.55**	.57**	.62**	—										
5. Moral obligation ^b	1.48	3.16	-.56**	.54**	.58**	.47**	—									
6. Behavioral beliefs ^b	0.52	104.24	-.47**	.38**	.57**	.42**	.41**	—								
7. Normative beliefs	-5.06	5.34	.52**	-.51**	-.51**	-.54**	-.56**	-.40**	—							
8. Alternative opportunities	5.96	1.49	.14	-.10	-.15*	-.14*	-.20**	-.31**	.20**	—						
9. Tenure	5.03	5.86	-.28*	.20**	.23**	.14*	.19**	.27**	-.20**	-.24**	—					
10. Marital status	1.59	0.82	-.03	-.04	-.04	-.03	-.02	-.11	.03	.02	-.08	—				
11. Educational level	2.11	0.88	.13	-.06	-.12	.01	-.10	-.23**	.11	.11	-.31**	.02	—			
12. Hospital location	1.86	0.34	-.07	.09	.08	.08	.11	.05	-.05	-.10	-.03	.04	.07	—		
13. Hospital size	2.64	0.65	.00	.03	-.01	.02	.04	.03	.01	-.10	-.02	.07	.07	.74**	—	
14. Full time versus part time	1.22	0.48	.17	-.06	-.04	-.12	-.05	-.01	.10	.00	-.05	-.14*	-.02	.06	-.01	—

^a Correlations with employment status are biserial correlations.
^b Differential scores were used.
* = *p* < .01. ** = *p* < .001.

differential measures for attitude ($\beta = .72, p < .001$), subjective norm ($\beta = .10, p < .01$), and moral obligation ($\beta = .10, p < .01$) were the only significant predictors of differential intention.

The holdback sample was then used to replicate the model developed on the initial sample. For the replication, the hierarchical regression analyses on the holdback sample indicated that the combination of all variables effectively predicted employment status ($R^2 = .35$). Again, differential intention was the only significant predictor; it accounted for 32% of the variance for employment status. Differential intention was also predicted ($R^2 = .58$) by the same three variables, differential attitude ($\beta = .60, p < .001$), differential subjective norm ($\beta = .15, p < .001$), and differential moral obligation ($\beta = .11, p < .01$).

The initial and holdback samples were combined in order to obtain a final set of regression weights. The hierarchical regression analyses on this sample indicated that the combination of all predictors explained 32% of the variance in employment status and 63% of the variance in differential intention. Once again, differential intention was the only significant predictor of status, and attitude ($\beta = .67, p < .001$), subjective norm ($\beta = .13, p < .001$), and moral obligation ($\beta = .07, p < .001$) were the only significant predictors of differential intention.

Comparison of Differentials Versus Resign-Only Scores

To determine the benefits of using differential measures rather than measures of a single behavioral alternative, the re-

Table 2
Hierarchical Regression Results for Employment Status and Intention for Each Sample

Predictor	Initial sample				Holdback sample				Combined sample			
	Employment status		Intention		Employment status		Intention		Employment status		Intention	
	β	ΔR^2	β	ΔR^2	β	ΔR^2	β	ΔR^2	β	ΔR^2	β	ΔR^2
Intention ^a	-.38**	.29			-.57**	.32			-.46**	.30		
Attitude ^a	-.08	.01	.72**	.65	-.10		.60**	.53	.01		.67**	.60
Subjective norm ^a	-.01		.10*	.02	-.16	.01	.15**	.03	-.10		.13**	.02
Moral obligation ^a	-.05		.10*	.01	-.06		.11*	.01	-.05		.07*	.01
Behavioral beliefs ^a	-.09		-.09		.00		.03		-.05		-.06	
Normative beliefs	.02		-.02		-.01		-.10		-.04		-.06	
Alternative opportunities	-.01		.03		-.01		.03		.00		.03	
Tenure	-.05		.03		.02		-.01		-.05		.01	
Marital status	-.03		-.02		-.07		.06		-.05		.01	
Educational level	.01		.03		.00		-.05		.02		-.01	
Hospital location	-.03		-.02		-.05		-.01		-.04		-.01	
Hospital size	.04		.05		.04		-.03		.04		.01	
Full time versus part time	.06		-.01		-.09		-.01		.00		-.04	
<i>R</i> ²	.32		.68		.35		.58		.32		.63	

^a Differential scores were used.
* = *p* < .01. ** = *p* < .001.

Table 3
Results of Factor Analysis of Belief Scores

Item	Factor loadings						Final communality estimate
	1	2	3	4	5	6	
Factor 1: Work Environment							
Can rely on the support of nursing administration	.78	.14	.14	.08	.17	.22	.74
Goals and concerns shared by hospital administration	.77	.21	.08	.14	.14	.21	.73
Work in an environment where there is a spirit of cooperation and teamwork	.61	.22	.41	.20	.06	.13	.66
Have autonomy and authority to use my judgment and make decisions about patient care	.61	.36	.24	.11	.06	.09	.59
Work in an environment where doctors treat me as a professional—with courtesy and respect	.58	.34	.17	.08	.10	.03	.51
Provides educational and learning experiences which enhance my professional growth	.60	.45	.11	.18	.04	.12	.63
Like and respect my co-workers	.55	.31	.45	.10	.00	.04	.62
Have a considerate and responsive supervisor—one who listens, understands my problems, and appreciates my work	.52	.09	.36	.21	.08	.32	.57
Factor 2: Nursing Practice							
Can use my nursing skills and keep them up to date	.24	.75	.14	.03	−.08	.05	.65
Working with the kind of patients I prefer	.26	.74	.17	.03	.03	.00	.65
Have opportunities for a variety of patient care experiences	.35	.63	.14	.13	−.16	−.08	.58
Job provides a sense of worth and a feeling of accomplishment	.26	.64	.38	.15	.20	.25	.76
Feel that I am helping people	.07	.59	.54	.09	.11	.12	.69
Have a challenging, stimulating, and interesting job	.21	.59	.42	.13	.11	.26	.68
Can provide my patients with more than just physical care (e.g., teaching, emotional and family support, follow-up)	.55	.49	−.01	.16	.17	−.08	.61
Can provide the quality of care patients deserve	.48	.48	.15	.16	.35	−.02	.63
Factor 3: Hygiene Factors							
Receive an acceptable salary	.17	.22	.70	.07	.10	.12	.60
Have job security	.43	.18	.59	.10	.11	.04	.58
Work on a schedule and/or shift I prefer	.11	.25	.56	.41	.14	.17	.61
Receive fringe benefits (e.g., insurance, retirement, sick leave)	.50	.10	.50	.22	.08	.07	.56
Factor 4: Opportunities Available by Resigning							
Have time for myself—to do things I enjoy	.16	.02	.17	.86	.23	−.03	.85
Have time for my family	.18	.04	.20	.84	.23	−.08	.84
Can meet and be with people	.16	.28	.07	.57	−.11	.40	.62
Advance my career	.27	.35	.04	.44	−.05	.31	.50
Factor 5: Physical–Emotional Costs							
Feel overworked and have too much to do	.15	.00	.13	.15	.81	.14	.74
Job is stressful and fatiguing	.15	.03	.11	.13	.80	.06	.71
Factor 6: Negative Job Characteristics							
Work in a setting where I am unfamiliar with the routine, equipment, and personnel	.02	.00	.21	.03	.04	.66	.47
Work is affected by poor communication and coordination between units or departments	.35	−.03	−.07	−.10	.25	.58	.54
Feel bored and restless	.11	.35	.06	.36	.06	.49	.53
Percentage of common variance	17	15	10	9	7	6	

gression analyses with all 13 predictors were repeated using measures of the theoretical components that related only to resigning behavior. At each step, the *R* resulting from this resign-only model was compared to the *R* for an expanded model in which the differential measures were entered into the equation. This expanded model (*R* = .33) accounted for significantly more variance in employment status, *F*(5, 407) = 5.47, *p* <

.001, than did the resign-only model (*R* = .29). Similarly, adding differential scores to the resign-only model significantly increased the *R* when predicting both intention to resign and differential intention. The *R* for the expanded model predicting intention to resign was .65, whereas the *R* for the resign-only model was .60, *F*(4, 412) = 14.64, *p* < .001. When predicting differential intention, the *R*(.68) for the expanded model was

Table 4
T Tests Comparing Stayers and Leavers on the Belief Factors

Factor	M		<i>t</i> ^a	<i>p</i>
	Stayers	Leavers		
Work Environment	1.38	−24.68	5.36	.001
Nursing Practice	8.30	−19.39	6.25	.001
Hygiene Factors	3.72	−7.15	4.17	.001
Opportunities Available by Resigning	0.31	−20.53	6.40	.001
Physical–Emotional Costs	−2.61	−5.27	1.88	<i>ns</i>
Negative Job Characteristics	2.42	−1.59	2.92	.01

^a *df* = 439 for all *t* tests.

significantly greater, $F(4, 412) = 25.81, p < .001$, than the $R(.60)$ for the resign-only model. Finally, the correlation between differential attitude ($r = .57$) was a significant improvement, $t(438) = 25.64, p < .001$, over the correlation between attitude and the behavioral beliefs for resigning behavior ($r = .21$). Replicating these analyses on the holdback and combined samples produced identical results. These consistent and significant results would appear to justify using differential measures in the development of the turnover model, and in addition, differential measures were used in subsequent analyses.

Belief Data Analyses

Because the best predictor of intention was attitude, it is important for both theoretical and applied reasons, to identify the specific sources of the attitude. According to the theory, the attitude toward an act should be a function of the sum of the behavioral beliefs. The results indicate that differential attitude and the sum of the 29 differential beliefs are significantly correlated for the initial sample ($r = .57, p < .001$) and for the holdback sample ($r = .67, p < .001$). To summarize the 29 differential beliefs, a principal components analysis using varimax rotations was applied to the belief items. Six factors were extracted: (a) Work Environment, (b) Nursing Practice, (c) Hygiene Factors, (d) Opportunities Available by Resigning, (e) Physical–Emotional Costs, and (f) Negative Job Characteristics. These six factors accounted for 64% of the variance in attitude. The 29 belief items and the factor loadings are presented in Table 3.

The sum of the items in each factor was determined for each nurse. (The means for stayers and leavers for these factors are contained in Table 4.) A *t* test was performed for each factor, comparing the stayers with the leavers. The results of these *t* tests indicated that there were significant differences for five of the six factors. These factors are Work Environment, Nursing Practice, Hygiene Factors, Opportunities Available by Resigning, and Negative Job Characteristics. Only the Physical–Emotional Costs (burnout) factor revealed no significant differences between stayers and leavers. These results are summarized in Table 4.

Discussion

The results of this study provide substantial support for applying the theory of reasoned action to the turnover area. The

resulting model accounted for 32% of the variance when predicting turnover. Consistent with the theory, turnover was determined by the nurses’ differential intention. This finding is in agreement with previous applications of the theory (Newman, 1974; Hom & Hulin, 1981), as well as with a recent meta-analysis of employee turnover (Steel & Ovalle, 1984).

The nurses’ differential intentions, in turn, were predicted by three differentially measured components: their attitude toward the act, their perception of the social influence to remain or resign, and their feelings of moral obligation. The combination of predictors accounted for 68% of the variance in intention.

Furthermore, consistent with the recommendations of Mobley et al. (1979), this model was subjected to a replication procedure using a holdback sample. This replication model produced the same set of predictors.

It was the purpose of this research not only to apply the latest version of the theory of reasoned action in a typical turnover situation for the first time, but also to make refinements in the theory and to evaluate the usefulness of some recent methodological suggestions for improving the usefulness of the theory. Therefore, this research evaluated the effectiveness of the theoretical construct of moral obligation and of using differential scores versus resign-only scores.

With respect to moral obligation, a variable not included in the theory of reasoned action, the results indicated that differential moral obligation was a significant predictor of differential intention. This result is in contrast to that of Hom and Hulin (1981), who found that moral obligation did not add to the theory of reasoned action’s predictive power with respect to national guardsmen’s reenlistment decisions. This suggests that the role of moral obligation in turnover decisions depends on the occupation and the subgroup being investigated.

The comparisons of the resign-only model and the expanded model consistently demonstrated the advantage of using measures that relate to both behavioral alternatives. At all levels in the sequences, differential measures made a significant contribution to the variance accounted for by resign-only items. This improvement was especially dramatic when predicting attitudes. These results support the contention of Fishbein (1980) and suggest that the predictive validity of future models of turnover would be enhanced by considering all of the available decision alternatives. These results should be interpreted with some caution, however, because it is possible that they may reflect increased reliability resulting from using two items rather than one item to measure a construct.

The present research also indicated that the theory of reasoned action has the potential to be effectively applied in hospital organizations to reduce nurse turnover. Hospital administrators and nursing directors armed with a list of the belief statements will have an understanding of some of the underlying causes of a nurse’s attitudes, intentions, and behavior with respect to turnover. These results, as well as those from the entire model (e.g., importance of moral obligation), can be used as a basis for designing intervention programs to reduce both nurse turnover and the intention to turnover.

In summary, it seems clear that the theory of reasoned action has demonstrated its potential usefulness in the turnover area. Like all theoretical approaches in I/O, it awaits further theoretical and methodological refinements, like those investigated in this research, to improve its predictive effectiveness.

References

- Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood Cliffs, NJ: Prentice-Hall.
- Cotton, J. L., & Tuttle, J. M. (1986). Employee turnover: A meta-analysis and review with implications for research. *Academy of Management Review*, 11, 55-70.
- Fishbein, M. (1967). Attitude and the prediction of behavior. In M. Fishbein (Ed.), *Readings in attitude theory and measurement* (pp. 477-492). New York: Wiley.
- Fishbein, M. (1980). A theory of reasoned action: Some applications and implications. In H. Howe & M. Page (Eds.), *Nebraska Symposium on Motivation*, (Vol. 27, pp. 65-116). Lincoln: University of Nebraska Press.
- Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior*. Reading, MA: Addison-Wesley.
- Hom, P. W., Griffeth, R. W., & Sellaro, C. L. (1984). The validity of Mobley's (1977) model of employee turnover. *Organizational Behavior and Human Performance*, 34, 141-174.
- Hom, P. W., & Hulin, C. L. (1981). A competitive test of the prediction of reenlistment by several models. *Journal of Applied Psychology*, 66, 23-39.
- Hulin, C., Roznowski, M., & Hachiya, P. M. (1985). Alternative opportunities and withdrawal decisions: Empirical and theoretical discrepancies and an integration. *Psychological Bulletin*, 97, 233-250.
- Hwalek, M., Firestone, I., & Hoffman, W. (1982). The role social pressures play in early retirement propensities. *Aging and Work*, 5, 157-168.
- Kramer, M. (1981). Philosophical foundations of baccalaureate nursing education. *Nursing Outlook*, 4, 224-228.
- Kunin, T. (1955). The construction of a new type of attitude measure. *Personnel Psychology*, 8, 65-77.
- Mobley, W. (1977). Intermediate linkages in the relationship between job satisfaction and employee turnover. *Journal of Applied Psychology*, 62, 237-240.
- Mobley, W., Griffeth, R., Hand, H., & Meglino, B. (1979). Review and conceptual analysis of the employee turnover process. *Psychological Bulletin*, 86, 493-522.
- Newman, J. (1974). Predicting absenteeism and turnover: A field comparison of Fishbein's model and traditional job attitude measures. *Journal of Applied Psychology*, 59, 610-615.
- Pomazal, R. J., & Jaccard, J. J. (1976). An informational approach to altruistic behavior. *Journal of Personality and Social Psychology*, 33, 317-326.
- Porter, L., & Steers, R. (1973). Organizational, work, and personal factors in employee turnover and absenteeism. *Psychological Bulletin*, 80, 151-176.
- Porter, L., Steers, R., Mowday, R., & Boulian, P. (1974). Organizational commitment, job satisfaction, and turnover among psychiatric technicians. *Journal of Applied Psychology*, 59, 603-609.
- Price, J. L., & Mueller, C. W. (1981). A causal model of turnover for nurses. *Academy of Management Journal*, 24, 543-565.
- Seybolt, J., Pavett, C., & Walker, D. (1978). Turnover among nurses: It can be managed. *Journal of Nursing Administration*, 8, 4-9.
- Sheridan, J. E., & Abelson, M. A. (1983). Cusp catastrophe model of employee turnover. *Academy of Management Journal*, 3, 418-436.
- Smetana, J., & Adler, N. (1980). Fishbein's Value \times Expectancy model. *Personality and Social Psychology Bulletin*, 6, 89-96.
- Smith, P., Kendall, L., & Hulin, C. (1969). *The measurement of satisfaction in work and retirement*. Chicago: Rand McNally.
- Steel, R. P., & Ovalle, N. K., 2d. (1984). A review and meta-analysis of research on the relationship between behavioral intentions and employee turnover. *Journal of Applied Psychology*, 69, 673-686.
- Triandis, H. (1975). Cultural training, cognitive complexity, and interpersonal attitudes. In R. Brislin, S. Bochner, & W. Lonner (Eds.), *Cross-cultural perspectives on learning* (pp. 39-77). New York: Sage/Halsted/Wiley.
- Wandelt, M. A., Pierce, P. M., & Widdowson, R. R. (1981). Why nurses leave nursing and what can be done about it. *American Journal of Nursing*, 1, 73-77.
- Wolf, G. A. (1981). Nursing turnover: Some causes and solutions. *Nursing Outlook*, 4, 233-236.

Received March 17, 1986

Revision received September 26, 1986

Accepted December 1, 1986 ■

Power Relinquishment Versus Power Sharing: Theoretical Clarification and Empirical Comparison of Delegation and Participation

Carrie R. Leana

Department of Management, University of Florida

This article presents a theoretical and empirical comparison of delegation and participation. Although the two processes have sometimes been treated as interchangeable, delegation and participation have evolved from two different theoretical perspectives and are used by managers under different sets of conditions. Two studies are reported that examined these differences. The experimental study examined situational factors in Vroom and Yetton's (1973) leadership model that predict differences in managers' reported preferences for delegation or participation. Results indicated that decision importance, subordinate information, and subordinate goal congruence explained 23% of the variance in managers' preferences. The correlational study examined similar situational predictors of supervisors' reported use of delegation and participation with subordinates. These results largely confirmed the findings of the experimental study and also showed supervisor workload as a significant predictor. In addition, objective measures of subordinate performance significantly correlated with the use of delegation but not with participation. The implications of the findings for research on participative decision making are discussed.

Research on the distribution of decision-making authority in organizations has been both abundant and diverse. This literature spans a broad range of research topics, from participation in decision making to worker autonomy to structural decentralization. The largest body of literature addressing this issue, however, is the research investigating subordinate participation in decision making (PDM). This research is itself quite diverse and plagued with inconsistencies concerning both the definition and the implementation of participative decision-making processes (Schweiger & Leana, 1986). Participation can vary in scope, content, and degree, in whether it is formal or informal, and in whether it is forced or voluntary (Locke & Schweiger, 1979). Moreover, PDM can take many different forms ranging from subordinate consultation to superior-subordinate decision making to participation through subordinate representation.

Although these variations in method suggest that little agreement exists on the exact meaning of PDM, participation has commonly been operationally defined by researchers as joint decision making between superior and subordinate (Bass, 1981). Consequently, much of the PDM research has tended to focus exclusively on comparisons between joint decision making and autocratic arrangements in which subordinates are not included in any aspect of the decision-making process. Conversely, research involving comparisons of other methods of in-

volving subordinates in decision making is decidedly rare. One such method that has suffered from a particular lack of recognition by researchers is *delegation*.

Delegation is a process whereby the manager transfers decision-making authority to a subordinate. Although delegation has received considerable attention in the management literature, this attention has been limited primarily to nonempirical books and journals (e.g., McConkey, 1974). Researchers have largely ignored delegation, preferring instead to focus on participation with its corresponding social and ideological connotations (Locke & Schweiger, 1979). In the process, both the conceptual and the operational differences between participation (i.e., joint decision making) and delegation have tended to become obscured.

As Strauss (1963), Heller (1976), and Locke and Schweiger (1979) have suggested, participative and delegative means of distributing decision-making authority are conceptually and practically quite different. Participation finds its theoretical roots in a human relations approach to management that emphasizes power equalization and social interaction. Likert (1967) and other proponents of a human relations model specifically exclude delegation as a viable alternative to participation because, as Heller (1976) has suggested, "they see a special virtue in group interaction which they cannot extend to a relatively solitary activity" (p. 701). Conversely, delegation finds its proponents in cognitive growth or human resource approaches that emphasize the need for subordinate autonomy and individual development. These approaches are manifested in various theories of human motivation ranging from Maslow's (1943) discussion of worker needs for self-actualization, to more recent developments in the design of work (e.g., Hackman & Oldham, 1976). The theoretical rationale for delegation, then, is quite different from participation in that it focuses on developing individual autonomy rather than on engendering democratic (i.e.,

Parts of this research were based on a doctoral dissertation completed at the University of Houston in 1984.

The author would like to thank Art Jago for providing data reported in parts of the research, and David Schweiger and Chet Schriesheim for their useful comments on earlier drafts of the article.

Correspondence concerning this article should be addressed to Carrie R. Leana, who is now at the Graduate School of Business, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

participative) processes. As Strauss (1963) suggested, "In a sense, participation emphasizes equality and conformity while delegation encourages achievement and individualism" (p. 43). Thus, from a motivational perspective, participation and delegation are geared toward attaining two very different ends.

Participation and delegation also differ with regard to the means with which they attempt to attain their respective ends, both in the processes used and in the amount or degree of authority actually vested in subordinates. Although participation entails superior-subordinate collaboration in decision making, delegation refers to the actual passing of decision-making authority from superior to subordinate (Locke & Schweiger, 1979; Bass, 1981). In this regard, participation is always a dyadic or group decision-making process, whereas delegation typically entails decision making by an individual. Moreover, delegation necessarily entails assigning authority for decision choice to the subordinate, whereas participative methods do not involve subordinate autonomy in decision choice (Heller & Yukl, 1969; Strauss, 1963; Vroom & Yetton, 1973). This latter point is a fundamental distinction between the two processes. As Locke and Schweiger (1979) suggested, "the result [of delegation] is not a 'sharing in common' with others but rather a specific division of labor which is determined hierarchically" (p. 274).

This distinction is of particular importance because it suggests that despite many researchers' almost exclusive focus on participation, delegation may be far more central to the issue of organizational distributions of authority. Through delegation, subordinates are given autonomy in making organizational decisions and are vested with authority to make choices regarding decision outcomes. Conversely, participation does not entail this autonomy. Instead, it implies that decision-making authority will be shared between superior and subordinate and does not entail the passing of authority for decision choice to a lower organizational level. In this regard, delegation has correctly been conceptualized as the more complete form of subordinate involvement in decision making (e.g., Heller & Yukl, 1969; Tannenbaum & Schmidt, 1958; Vroom & Yetton, 1973) because, unlike participation, it necessarily entails a relinquishing of managerial control or power over decision choice.

Despite the managerial relinquishing of control entailed in delegation, Bass and Valenzi (1974) found participation to be only slightly more prevalent than delegation in a broad sample of organizations. In addition, participation and delegation were both reported as significantly more prevalent than autocratic decision making. This latter finding may partially explain why so much research has concentrated on participation; it does not, however, explain why researchers have largely excluded delegation from their studies. Although many managers may delegate decision making nearly as much as they share it (Bass & Valenzi, 1974), researchers have had little to say about delegation. Instead, most of our knowledge of delegation is derived from anecdotal accounts by practicing managers and consultants (e.g., McConkey, 1974; Steinmetz, 1976). Moreover, researchers who have included delegation in their analyses have largely treated it as a next-step progression that differs from participation only in degree and not in any substantive theoretical or operational form (cf. Bass & Valenzi, 1974; Heller, 1976; Tannenbaum & Schmidt, 1958; Vroom & Yetton, 1973). Thus, despite the paucity of research on delegation and the rather sub-

stantial theoretical and operational distinctions that can be made between delegation and participation, researchers have either ignored these distinctions or, at best, treated the two processes as progressions on the same participation continuum.

This article is an exploratory attempt to examine the similarities and differences between participation and delegation. Two studies are reported. The first is a reanalysis of data testing the Vroom and Yetton (1973) leadership model that was originally reported by Vroom and Jago (1974). In this study the effects of seven situational predictors on managers' selections of delegation versus participation (joint decision making) were examined. In the second study data were gathered in a field setting, and predictors of managers' perceived use of delegation versus participation were examined. The relationship of subordinate performance and satisfaction with these two decision-making processes were also examined in this study.

Study 1: Experimental Examination of Predictors of Delegation and Participation

Method

The data used in this analysis, originally reported by Vroom and Jago (1974), consisted of responses to 24 standardized cases by 98 managers from a variety of business and government organizations. In responding to the standardized cases, participants were asked to assume the role of leader and indicate which of five decision processes they would use in each of the 24 situations described. The alternative decision processes ranged from autocratic decision making (AI and AII) to subordinate consultation (CI) to joint decision making between manager and subordinate (GI) to delegation (DI). Consistent with the purposes of this article, the reanalysis reported here focused exclusively on distinctions made by respondents between GI and DI. In total, there were 2,352 responses (24 cases completed by 98 respondents), although only 1,002 of these were GI or DI responses and thus included in the analysis.

The actual cases that were used by Vroom and Jago (1974) incorporated combinations of seven situational attributes that were hypothesized to influence the appropriateness of the decision process used. These seven attributes were (a) the importance ("quality requirement") of the decision, (b) the extent to which the leader possessed sufficient information or expertise to make a high quality decision him or herself, (c) the extent to which subordinates had all of the information necessary to make a good decision, (d) the extent to which the problem was structured, (e) the extent to which acceptance or commitment on the part of subordinates was critical to the effective implementation of the decision, (f) the prior probability that the leader's autocratic decision would be accepted by subordinates, and (g) the extent to which subordinates were motivated to attain the organizational goals or objectives explicit in the problem. The various situational attributes were manipulated as either present or absent within the 24 cases completed by each respondent. Examples of the actual cases used and further explanation of how the cases were constructed is provided by Vroom and Yetton (1973) and Vroom and Jago (1974).

Results

The reanalysis focused exclusively on distinctions made by respondents between participation (GI) and delegation (DI). In this reanalysis, the situational attributes contained in the cases were entered in a regression equation using a hierarchical procedure. This hierarchical format was derived from Vroom and Yetton's (1973) decision tree, which specifies the circumstances

Table 1
*Hierarchical Regression Analysis of Vroom and Yetton Problem Attributes as Predictors
 of Managers' Choices of Delegation and Participation (Study 1)*

Problem attribute	Decision process			
	β^a	R^2 change	Total R^2	F
Step 1				
Decision importance (quality)	-.337	.11	.11	128.34**
Step 2				
Subordinate information	.307	.07	.18	90.81**
Subordinate goal congruence	.228	.05	.23	62.57**
Step 3				
Problem structure	.062	.00	.23	3.91*
Prior probability of acceptance	.118	.01	.24	17.39**
Leader information	.033	.00	.24	0.56
Importance of acceptance	-.038	.00	.24	1.44

Note. $N = 1,002$.

^a Positive betas indicate greater use of delegation; negative betas indicate greater use of participation.

* $p < .05$. ** $p < .01$.

under which either of the two processes is most likely to be used. Using this framework, the importance of the decision (attribute a) was entered first in the equation, followed in the next step by the two subordinate characteristics (attributes c and g). At the third step, the other four problem attributes (attributes b, d, e, and f) were entered into the regression equation. The decision tree itself is based on a complex set of rules governing the appropriateness of each decision process under different combinations of the situational attributes. The reader is referred to Vroom and Yetton (1973) and Vroom and Jago (1974) for a complete discussion of the rules underlying the decision tree.

Table 1 shows the results of the hierarchical regression. As indicated in the table, the importance of the decision (attribute a) explained 11% of the variance in managers' selections of joint decision making over delegation. The second set of variables entered in the equation, subordinate information (attribute c) and subordinate goal congruence (attribute g), respectively explained an additional 5% and 7% of the variance in managers' choices of decision-making processes. Four predictors were entered in the last step of the regression. Two of these, problem structure (attribute d) and prior probability of acceptance (attribute f), were statistically significant, although neither explained more than 1% of the variance in managers' choices and, for this reason, were judged to be weak predictors of the differences between delegation and participation.

The beta weights reported in Table 1 indicate that the importance of the decision (attribute a) was negatively related to managers' choices of delegation. That is, managers chose to share (GI) rather than delegate (DI) decisions that were described as consequential to the organization. For decisions described as relatively inconsequential, however, managers indicated that they would use delegation. These results are consistent with Heller's (1971, 1973) survey research in which managers reported delegating organizationally important decisions less than 5% of the time.

Subordinate information and subordinate goal congruence together explained 12% of the variance in managers' selections of decision-making processes. Managers indicated that they

would relinquish (DI) rather than share (GI) decision-making authority when the problem scenarios described the subordinate as having sufficient information to make the decision and as sharing organizational goals. Thus, both characteristics of the subordinate substantially influenced managers' reported willingness to delegate rather than share decision-making authority.

Study 2: Correlational Examination of Predictors and Consequences of Delegation and Participation

Method

In this study, 26 insurance claims supervisors reported on their use of delegation and participation (joint decision making) in overseeing the work of 122 claims adjusters. The supervisors and adjusters worked in 19 branch offices of a large national insurance company. The claims adjusters were responsible for assessing the dollar value of a variety of insurance claims ranging from automobile and real property damage to bodily injury and personal liability settlements. Each supervisor managed from 3 to 8 claims adjusters and worked closely with them on calculating insurance settlements.

Data were gathered from questionnaires completed by the supervisors and the claims adjusters, and from archival information drawn from company records. Supervisors were asked to report the percentage of time they used each of four decision processes with each adjuster. These decision processes were (a) autocratic decision making, (b) consultation, (c) participation, and (d) delegation. The descriptions of the four processes were identical to those provided to respondents in the Vroom and Jago (1974) experimental study (Study 1) previously described. Like Study 1, this analysis focused exclusively on distinctions made between participation and delegation.

Predictors of decision processes. The three significant predictors in the previous examination of delegation and participation (Study 1) were also examined in this study. However, the present analysis focused on differences among subordinates rather than on differences in various decisions faced by the managers. The three similar predictors were (a) the importance of the decisions made by each subordinate, (b) each subordinate's job capability (similar to subordinate information in Study 1), and (c) each subordinate's trustworthiness (similar to subordinate goal congruence in Study 1). A fourth predictor, the work load

managed by each supervisor, was also assessed. This was done to examine the possible effects of work-load pressure on a supervisor's use of delegation, as suggested by Yukl (1981) and in numerous practitioner accounts of "effective" delegation (e.g., McConkey, 1974; Steinmetz, 1976).

The importance of decisions made by a subordinate was assessed by examining the types of claims each adjuster was responsible for processing. These claims types ranged from automobile property damage to real property damage to personal injury and liability damage. The importance of each of the types of claims was consensually determined by three regional managers who oversaw the branch claims offices within their respective jurisdictions. These regional managers ranked personal liability claims to be most important, real property damage claims next most important, and automobile property damage claims least important. Each adjuster specialized in one of these claims types, handling, for example, automobile property damages exclusively.

A claims adjuster's job capability and trustworthiness were assessed by asking his or her supervisor to rate the adjuster on these dimensions. Each supervisor was asked to position the initials of all of his or her subordinates on each of six 15-c lines. Three of these measured perceived job capability and three measured perceived trustworthiness. The six lines (e.g., "Understanding of job requirements") were anchored at either end (e.g., *Thorough understanding vs. Incomplete understanding*), and the items were scored from 1 through 15 according to their proximity to the closest centimeter. Reliability estimates (coefficient alpha) were .91 for the three capability items and .71 for the three trustworthiness items.

Supervisor work load was measured by the volume of claims managed by each supervisor for the 6-month period prior to data collection. The mean number of claims managed by each supervisor was 1,452 ($SD = 1,002$). Data on claims adjuster demographics and job tenure were also gathered through questionnaire items completed by the claims adjusters.

Consequences of decision processes. Subordinate job performance and satisfaction were assessed as potential outcomes of participation and delegation. Satisfaction with supervision and global job satisfaction were measured using items from the Job Diagnostic Survey (Hackman & Oldham, 1975) that were completed by the claims adjusters. Reliability estimates (coefficient alpha) for these scales were .92 and .77, respectively. Job performance was assessed by examining (a) the number of claims processed by each adjuster over a 3-month period, and (b) the average cost of claims that were settled by each adjuster within the same time period. Note that with this latter measure, lower costs indicated better job performance.

Results

Table 2 shows the correlations between the predictors and consequences previously described and supervisors' reports of their use of delegation and participation. As indicated in the table, the correlations associated with the reported use of participation and delegation were quite different. Like the results in Study 1, the importance of decisions made by subordinates was negatively related to reported delegation and positively related to participation. Moreover, supervisor perceptions of subordinate job capability (similar to subordinate information in Study 1) and trustworthiness (similar to goal congruence) were positively related to supervisors' reported use of delegation. Subordinate job capability was negatively related to supervisors' reported use of participation, although trustworthiness was not significantly correlated with participation.

Other differences were apparent in the relationships among situational factors and supervisors' reports of their relative use

Table 2

Correlates of the Reported Use of Delegation and Participation by Insurance Claims Supervisors (Study 2)

Variable	Participation	Delegation
Predictors		
Decision importance	-.18*	-.16*
Perceived subordinate job capability	-.16*	.42**
Perceived subordinate trustworthiness	.07	.27**
Supervisor work load	-.32**	.28**
Subordinate demographic		
Age	-.23**	.17*
Sex ^a	.30**	-.22**
Job tenure	-.24**	.19*
Consequences		
Subordinate satisfaction		
With job	-.04	.01
With supervisor	.07	-.02
Subordinate performance		
Claims settled	-.29**	.30**
Cost of claims	.10	-.15*

Note. $N = 122$.

^a 1 = male; 2 = female.

* $p < .05$. ** $p < .01$.

of each process. Most notably, supervisor work load was positively associated with the reported use of delegation and negatively associated with the reported use of participation. Similarly, the correlations regarding subordinate demographic characteristics shown in Table 2 indicated differences between the two processes. Supervisors reported using delegation for subordinates who were older, had more job experience, and were men. Conversely, the reported use of participation was more prevalent with claims adjusters who were younger, had less job experience, and were women.

Table 2 also reports the correlations between claims adjusters' satisfaction and performance and supervisors' reported use of the two processes. Neither global job satisfaction nor satisfaction with supervision was significantly correlated with delegation or participation. Delegation, however, was significantly correlated with both indicators of job performance. That is, those adjusters who were delegated more decision-making authority also tended to process more claims and at a lower average cost. Conversely, the use of participation was negatively correlated with the number of claims settled and was positively but not significantly correlated with average claim costs. When the effects of supervisor perceptions of subordinate capability were partialled from these correlations, the pattern of relations remained unchanged.

Table 3 shows the results of a regression analysis of the predictors of delegation and participation. This analysis followed a hierarchical format similar to that used in Study 1. Thus, the importance of the type of claims managed by each adjuster (decision importance) was entered first, followed by perceived subordinate characteristics (job capability and trustworthiness). A final situational factor, supervisor work load, was entered at the last step. The dependent variable, the relative use of delegation versus participation, was calculated for each subordinate by

Table 3
*Hierarchical Regression Analysis of the Relative Use of
 Delegation and Participation by Insurance
 Claims Supervisors (Study 2)*

Predictor	Decision process			
	β^a	R^2 change	Total R^2	F
Step 1				
Decision importance	-.193	.04	.04	4.62*
Step 2				
Perceived job capability	.426	.12	.16	8.13**
Perceived trustworthiness	.120	.00	.16	.00
Step 3				
Supervisor workload	.263	.06	.22	9.18**

Note. $N = 122$.

^a Positive betas indicate greater use of delegation; negative betas indicate greater use of participation.

* $p < .05$, ** $p < .01$.

subtracting the percentage of decisions that were reported as jointly made by the supervisor and subordinate from the percentage that were reported as delegated to the subordinate to make him or herself.

As indicated in Table 3, decision importance was a significant predictor of a supervisor's relative use of delegation versus participation. Perceived subordinate job capability was also a significant predictor. In essence, supervisors tended to delegate rather than share decision making with a claims adjuster who handled less important claims and who was perceived to be capable of performing his or her job. These results are quite consistent with those reported in Study 1.

Somewhat inconsistent with these previous results, however, was the inconsequential effect of perceived subordinate trustworthiness on supervisors' relative use of delegation and participation. As indicated in Table 3, perceived subordinate trustworthiness was not a significant predictor of a supervisor's relative use of each decision process. Conversely, in Study 1 subordinate trustworthiness (goal congruence) explained 5% of the variance in managers' choices of delegation or participation. This difference in results may be attributable to many factors (e.g., collinearity of trustworthiness with other predictors in Study 2), but one that seems most obvious concerns differences in how this variable was operationally defined in the two studies. In the Vroom and Yetton scenarios (Study 1), trustworthiness was assessed as supervisor-subordinate goal congruence with respect to the objectives of the specific decision described. For the insurance adjusters (Study 2), however, trustworthiness was assessed as a more global evaluation concerning overall congruence between supervisor and subordinate work objectives.

The final predictor examined in this analysis was supervisor work load. Consistent with the correlations reported in Table 2, a supervisor's work-load was a significant predictor of his or her relative use of delegation versus participation. Supervisors who had more work to manage tended to favor delegation over participation and thus relinquished authority rather than shared it with their subordinates. This finding is consistent with Yukl's (1981) prediction and much of the practitioner literature, which

has long touted delegation as a useful technique in time management.

Discussion

The results of the two studies revealed several interesting differences in the relative use of delegation and participation. First, the results indicated that managers do make distinctions between the two processes. Moreover, these distinctions were made in both the disposition of specific decisions (Study 1) and in a manager's overall dealings with particular subordinates (Study 2).

Second, supervisors chose to relinquish rather than share authority based on characteristics of the decision situation as well as characteristics of the involved subordinate. In both studies reported here, managers were willing to delegate control over decision making when a subordinate was viewed as capable of making good decisions and when the tasks or decisions in question were of lesser consequence to the organization. The results of Study 2 also indicated that work-load pressures may lead supervisors to delegate decisions over which they might otherwise have retained some control.

The effects of the supervisor's trust or confidence in subordinate judgment were inconsistent. In the experimental study (Study 1), goal congruence was a significant predictor of delegation. In the correlational study (Study 2), however, a supervisor's expressed trust in a subordinate did not significantly affect his or her relative use of delegation. As noted previously, this inconsistency in results may be attributable to differences in the level of specificity at which this variable was examined. Supervisors may be willing, for example, to delegate some decisions to a subordinate who can be trusted in particular aspects of a task, without having overall or general trust in the subordinate. Thus, goal congruence appears to be an important influence on managers' choices of decision-making processes as it is related to the particular decision to be made rather than the particular subordinate involved. That is, a subordinate may be trusted to make only some decisions him or herself, depending on the nature of the decision. Conversely, job capability did not appear to be decision specific and was a significant predictor in both studies.

These results have several implications for the literature on participative decision making. Most prominently, they call into question the assumption that delegation is merely a next-step progression on the participation continuum offered by many researchers (cf. Tannenbaum & Schmidt, 1958; Vroom & Yetton, 1973). Managers apparently view the relinquishing of control entailed in delegation as important enough to be undertaken only in the presence of a very limited set of circumstances (i.e., when less important decisions are made by more capable subordinates and when the supervisor is too overloaded with work to participate in the process). Moreover, in the correlational study, delegation was related to enhanced subordinate performance, whereas participation was associated with decreased performance by subordinates. These findings point to further potential differences between the two processes.

This is not to suggest that delegation and participation are entirely unrelated in both their antecedents and, perhaps more important, in their effects. Indeed, delegation, like shared deci-

sion making, is an indication of the distribution of authority in organizations. In this regard, it should share—and may in fact enhance—many of the presumed motivational benefits traditionally ascribed to participation, such as increased feelings of responsibility and ownership of work (Locke & Schweiger, 1979). It should be recognized, however, that the mechanisms governing delegation may not be the same as those governing participation. In their examination of different decision-making styles, Vroom and Jago (1974), for example, found a positive relationship between participativeness and decision importance for their *group* model in which shared decision making is described as the most participative process. However, in their examination of the *individual* model—which describes delegation as the most complete form of subordinate control—decision importance was negatively associated with the degree of subordinate involvement in decision making. Similarly, Bass, Valenzi, Farrow, and Solomon (1975) found delegation to be prevalent when subordinates engaged in complex work activities, whereas shared decision making was not significantly related to the complexity of subordinate tasks. Moreover, the results of this research indicated that the cognitive factors such as increased information and understanding that have traditionally been described as beneficial *outcomes* of participation (Locke & Schweiger, 1979), may in fact be required *antecedents* rather than consequences of delegation.

Thus, these results, as well as the work of previous investigators, point to the potential differences between delegation and participation that have been suggested by Strauss (1963) and others (e.g., Heller, 1976; Locke & Schweiger, 1979). Although this research was an initial attempt to explore these distinctions, several issues remain unresolved. First, neither study accounted for more than 24% of the variance in managers' selections of delegation or participation. Second, there were limitations to each of the studies. Both studies relied on managers' reports of their use of each process, rather than a more objective observational approach. In Study 2, for example, the relationships between the decision processes and the perceived subordinate characteristics may have been distorted due to the retrospective and self-report nature of the data. Moreover, neither study examined the two processes over time. Grove (1983) has suggested that delegation may be preceded by numerous joint decision-making sessions, perhaps to bolster the supervisor's confidence in the subordinate's ability. Also, the decisions being made by the subordinates in Study 2 were technical rather than managerial in nature. For this reason, the results may not be generalizable to other samples. Finally, neither study used longitudinal data.

Despite these limitations, this research can serve as an initial point from which future studies might be generated. The issue of authority distribution has been central in both historic and contemporary investigations of organizational behavior. From these results it seems clear that sharing authority and relinquishing it are very different processes for managers. These differences in both the antecedents and the specific outcomes

attributable to participation and delegation should be explored further in future studies.

References

- Bass, B. M. (1981). *Stogdill's handbook of leadership: A survey of theory and research*. New York: Free Press.
- Bass, B. M., & Valenzi, E. (1974). Contingent aspects of effective management styles. In J. G. Hunt & L. L. Larson (Eds.), *Contingency approaches to leadership* (pp. 130–152). Carbondale: Southern Illinois University Press.
- Bass, B. M., Valenzi, E., Farrow, D. L., & Solomon, R. J. (1975). Management styles associated with organizational, task, personal, and interpersonal contingencies. *Journal of Applied Psychology*, 60, 720–729.
- Grove, A. S. (1983). *High output management*. New York: Random House.
- Hackman, J. R., & Oldham, G. R. (1975). Development of the Job Diagnostic Survey. *Journal of Applied Psychology*, 60, 159–170.
- Hackman, J. R., & Oldham, G. R. (1976). Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance*, 16, 250–279.
- Heller, F. A. (1971). *Managerial decision making: A study of leadership styles and power sharing among senior managers*. London: Tavistock.
- Heller, F. A. (1973). Leadership, decision making and contingency theory. *Industrial Relations*, 12, 183–199.
- Heller, F. A. (1976). Decision-processes: An analysis of power-sharing at senior organizational levels. In R. Dubin (Ed.), *Handbook of work, organization, and society* (pp. 687–745). Chicago: Rand McNally.
- Heller, F. A., & Yukl, G. (1969). Participation, managerial decision making and situational variables. *Organizational Behavior and Human Performance*, 4, 227–241.
- Likert, R. (1967). *The human organization*. New York: McGraw-Hill.
- Locke, E. A., & Schweiger, D. M. (1979). Participation in decision-making: One more look. In B. M. Staw (Ed.), *Research in organizational behavior* (Vol. 1, pp. 265–340). Greenwich, CT: JAI Press.
- Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, 50, 370–396.
- McConkey, D. D. (1974). *No-nonsense delegation*. New York: AMACON.
- Schweiger, D. M., & Leana, C. R. (1986). Participation in decision making. In E. Locke (Ed.), *Generalizing from laboratory to field settings: Research findings from industrial-organizational psychology, organizational behavior, and human resource management* (pp. 147–166). Boston: D. C. Heath.
- Steinmetz, L. L. (1976). *The art and skill of delegation*. Reading, MA: Addison-Wesley.
- Strauss, G. (1963). Some notes on power-equalization. In H. Levitt (Ed.), *The social science of organizations* (pp. 40–84). Englewood Cliffs, NJ: Prentice-Hall.
- Tannenbaum, A., & Schmidt, W. (1958). How to choose a leadership pattern. *Harvard Business Review*, 36, 95–101.
- Vroom, V. H., & Jago, A. G. (1974). Decision making as a social process: Normative and descriptive models of leader behavior. *Decision Sciences*, 5, 743–769.
- Vroom, V. H., & Yetton, P. W. (1973). *Leadership and decision making*. Pittsburgh, PA: University of Pittsburgh Press.
- Yukl, G. (1981). *Leadership in organizations*. Englewood Cliffs, NJ: Prentice Hall.

Received March 26, 1986

Revision received October 20, 1986

Accepted October 20, 1986 ■

Judges' Mediation of Settlement Negotiations

James A. Wall, Jr. and Dale E. Rude
University of Missouri

This study investigates the effects that seven variables have on the level of judges' involvement in out-of-court settlement negotiations. In a nationwide survey, 650 judges and 1,100 attorneys were asked to read a civil case. The judges indicated how strongly they would mediate the negotiation between the opposing attorneys in the case and the specific tactics they would employ. The surveyed attorneys indicated the appropriate level of judicial involvement and the specific steps they felt a judge should take. The responses from the judges (58% response rate) and attorneys (73%) reveal that five variables—judges' and attorneys' attitudes toward judicial involvement, region of the country, respondent (judge versus attorney), size of case, and trial length—affect both the judges' mediation and the attorneys' preferences for judicial mediation. Two additional variables, the trial judge (current versus a different judge) and trier of the case (judge versus jury), had no effect on the dependent measures.

Over the years mediation has proved useful for resolving disputes in the arenas of international (e.g., Touval, 1982; Young, 1972), industrial (e.g., Kochan & Jick, 1978; Kolb, 1983), and marital (e.g., Haynes, 1981) relations. More recently, it has made significant inroads into the legal realm as judges have begun to mediate the out-of-court settlement of civil cases. This article reports the effects that seven variables have on the strength of judges' mediation and the tactics judges employ.

Settlement Negotiations

A civil dispute can be resolved in a trial, but most cases are "settled" out of court. Settlement is a negotiation process in which two persons, entities, or their attorneys attempt to hammer out an agreement. The prospective plaintiff usually initiates the process by hiring a lawyer who prepares a case against the defendant. Prior to filing the case, the plaintiff's attorney usually negotiates directly with the defendant or the defendant's attorney. If there is no agreement, the complaint is filed and the two attorneys may continue or enter into negotiations prior to a judge's involvement. If they are unsuccessful, the case is scheduled for a pretrial conference in which a judge reviews the case with the attorneys, rules on procedural issues, delineates which facts are or are not under agreement, sets a date for trial and, if he or she wishes, attempts to mediate an agreement. If there is no agreement, the case comes to trial, but even here the settlement negotiations can continue.

Out-of-court settlement yields many benefits to the disputing

parties (e.g., no costs of trial), to the judicial system (e.g., rapid disposition of cases), and to the judge (e.g., reduction of the case backlog); therefore, most judges—about 70% (Flanders, 1977)—mediate the settlement negotiations.

Judicial Mediation

In general, judges feel that their mediation techniques are effective (Wall, Schiller, & Ebert, 1984). Consequently, the judiciary in 1983 amended Federal Rule 16 to encourage judicial mediation in the federal courts. At the state level some courts (e.g., California) have introduced procedures such as week-long "trial holidays" in which all judges work at settling cases (Title, 1981). Many states also have appointed settlement judges whose major responsibility is case mediation.

As for practicing attorneys, most desire assistance from judges (Brazil, 1985) because they feel it expedites out-of-court settlement. Legal scholars agree (Galanter, 1985), adding that judicial involvement also improves the quality of the agreement, and recent evidence (Watson, 1984) indicates that judicial mediation increases the number of settlements.

Variables Affecting Judicial Mediation

What steps do judges take to mediate settlement and which variables influence their decision to mediate and their tactics in mediation? These questions, along with evidence that the dockets of civil courts are extensively backlogged, motivated us to undertake a concentrated study of judicial mediation.

In previous studies (Schiller & Wall, 1981) we identified 71 mediation techniques used by judges, which were analogous to those used by mediators in other arenas (Wall, 1981). We then determined how frequently each technique was employed (Wall & Schiller, 1982) and measured its perceived effectiveness (Wall, Schiller, & Ebert, 1984).

We then focused our attention on the variables that influence judges' mediation. In order to ferret out these variables, we interviewed 50 judges to find (a) which variables determine their

Support for this research was provided by the University of Missouri Graduate School and the Ponder Faculty Development and Research Fund.

The authors would like to thank Diane England for her assistance in this project. We are also indebted to Peter Carnevale, David Grigsby, and Dean Pruitt for their critiques of an earlier version of this article.

Correspondence concerning this article should be addressed to James A. Wall, Jr., Middlebush Hall, University of Missouri, Columbia, Missouri 65211.

involvement in the mediation of settlement negotiations and (b) the way in which the variables affect their involvement.

Some judges maintained that no situational variables altered their degree or manner of involvement. Several judges became very involved in all cases; other judges played a moderate role in every negotiation; and others participated in none. Most judges, however, indicated that their mediation efforts were situation dependent. About 20 variables were identified that apparently influence judicial efforts. This study examines the effects of 4 of these variables—case size, trial length, trier of case (Is the case to be bench- or jury-tried?), and trial judge (Is the case to be tried before the current judge or before another judge?). Two additional variables involving the influence of regional differences and the judges' attitude toward judicial involvement are also studied. The four principle variables and what the judges said about them follow.

Case size. The judges indicated that, in general, they wish to keep small cases out of court because their backlog is replete with such cases and that allowing more such cases into court would overwhelm the system. Therefore, they work diligently to mediate small cases to avoid tying up valuable court time. Because they feel that larger cases merit court time, they expend less effort to mediate them.

Trial length. The trial time required by the case also alters the judges' avowed mediation efforts. A case that promises to consume substantial trial time and to delay the hearing of subsequent cases, motivates judicial mediation. In this circumstance, judges realize that settling the case out of court reduces their trial docket and expedites the resolution of subsequent cases. Cases that will consume less time in trial provide less incentive for judges to mediate.

It should be noted here that case size and trial length vary independently in civil court cases. Cases can be large and promise to consume extensive trial time, or they can be small and brief. On the other hand, cases can be large and require a small amount of trial time (e.g., a liability suit in which a bus driver allegedly was careless and 10 people died), or they can be small and require extensive trial time (e.g., a suit in which a genetic engineering firm seeks to restrain a competing firm, by suing it for a modest amount, because the defendant allegedly violated a patent on a new bacterium).

Trier of the case and trial judge. These two variables interact to influence the judges' mediations. When a case is to be bench-tried before them, judges say that they are very reluctant to mediate the settlement negotiations because they feel that active mediation risks prejudicing themselves for the forthcoming trial. They mediate strongly, however, if the case is to be jury-tried or if another judge is to try the case. Specifically, if the case is to be bench-tried before another judge, the current judge feels his or her prejudices are of little consequence and, therefore, he or she feels free to mediate. Likewise, if the case is to be jury-tried before a different judge, the current judge feels free to intervene.

When the case is to be tried before the current judge with a jury deciding the case the judge mediates strongly. In such a situation, judges realize that their mediation endeavors and the responses of the attorneys might prejudice them toward the case, but, because the jury will decide the case, the judges believe the prejudice will not be deleterious.

Judges' attitudes toward mediation of settlement negotiations. As noted previously, many judges maintained, in their interview, that they harbored a predisposition toward mediating the settlement of cases. Their responses led to a hypothesis that each judge's attitude toward judicial mediation itself would affect how strongly he or she mediated the settlement of any particular case.

This hypothesis was supported by letters and phone calls from some judges in previous studies who voiced the opinion that judicial mediation of settlement negotiations was unethical, prejudiced, intolerable, and so forth, and therefore that they never undertook mediation. Those favoring judicial mediation proved equally communicative, arguing that they mediated in most cases and that judges who do not mediate are lazy, ill-informed, conservative, and complacent.

The hypothesis is consistent with Flanders' (1977), Carter's (1962), and Boyum's (1979) conclusions that a legal "culture" surrounds each judge and influences his or her settlement activities. They argue, in brief, that the environment of expectations, attitudes, and local practices surrounding a judge's court determines the judge's orientation or attitude toward settlement and, thereby, his or her involvement in mediation.

Sectional differences. Communications with judges and lawyers during previous studies support the assumption of cultural differences and indicate that there are important differences between geographic regions. The northeast and western states, with their large population centers and court loads, seem to possess norms favoring mediation of the settlement process. Thus, high levels of mediation are expected from judges in these areas. On the other hand, judges and attorneys in the southern and northcentral states seem less favorably disposed toward judicial involvement. Judges in these areas, therefore, mediate less strongly in cases coming before them.

Attorneys' Preferences

In recent years there has been a rising interest in attorneys' preferences for judicial mediation of settlement negotiations (Brazil, 1985). Our surveys and interviews have detected that, on occasion, attorneys wish for judges to involve themselves strongly in the settlement negotiations, and, at other times, they prefer for judges to eschew mediation. However, there is no current evidence as to which variables determine their preferences.

In order to overcome this deficiency, this study investigates the effects that the variables described in the preceding section have on lawyers' preferences for judicial mediation. The assumption underpinning the initial hypotheses is that attorneys have the same goals as the judges. That is, they want judges to expedite cases and to remain impartial. It is posited, given this assumption, that the variables under study—case size, trial length, trier of case, and trial judge—affect the attorneys' preferences for judicial mediation in a manner analogous to their effects on judges' mediation. Also it is expected that the attorneys' preference for judicial mediation in any particular case will correlate positively with their general attitude toward judicial mediation of settlement. Regional effects are also expected to be analogous to those found for judicial mediation.

Finally, a question arises about the relation between attorney preferences and actual judicial mediation. Do attorneys prefer

more or less mediation than judges give? Brazil's (1985) nationwide study of attorneys' views indicates that attorneys prefer more mediation. His findings reveal that, generally, attorneys would like judges to be more involved than they currently are in the settlement of civil cases. Consequently, we predicted that, in a particular case, attorneys prefer more judicial mediation than judges give.

Hypotheses

In sum, it is predicted that the variables discussed above—case size, trial length, trier of case, trial judge, attitudes toward judicial involvement in settlement, sectional differences, judges' preferences, and attorneys' preferences—have the following effects.

1. Judges' mediation efforts (and attorneys' preferences for judicial mediation) decrease as the size of the case increases.
2. The strength of judicial mediation efforts (and attorneys' preferences for judicial mediation) correspond to the trial time a case will consume.
3. The trial judge and trier of case variables interact. When a case is to be bench-tried before the current judge, he or she does not mediate strongly (and attorneys prefer that they do not become involved). In other cases (jury-tried, current judge; jury-tried, another judge; bench-tried, another judge), judges mediate strongly (and attorneys prefer that they become strongly involved).
4. The strength of the judges' mediation in a particular settlement negotiation is positively correlated with their attitude as to how strongly judges should, in general, facilitate the settlement of cases. Attorney preferences for judicial mediation also are positively correlated with their own general attitude.
5. Judges in the northeastern and western states mediate more often and forcefully than do their colleagues in northcentral and southern states. Likewise, attorneys in the northeastern and western states prefer stronger mediation than do those in the northcentral and southern states.
6. Attorneys prefer judicial mediation that is stronger than that offered by judges.

Method

Subjects

In order to test the hypotheses (described in the preceding section), cases and questionnaires were mailed to 650 state judges and 1,100 practicing attorneys chosen randomly from nationwide listings provided by the National Judicial College and the Association of Trial Lawyers of America.

Procedures

Each judge or attorney was sent a product liability case that we developed with the help of several judges. Before using the case, it was administered to 60 judges, who were asked if it was typical of civil cases brought before them. In their opinion it was a representative case. In the case, they were informed that the workers in a liquid fertilizer manufacturing plant were injured while using a toxic chemical and were, consequently, seeking damages from the producer of the chemical. The workers alleged that the producer was negligent in the production of the chemical and failed to provide proper instructions for its use. As a result

of that negligence, the workers (the plaintiffs) suffered dizziness, headaches, temporary loss of memory, and lengthy periods of fatigue.

The producer of the chemical (the defendant) argued that the chemical does not produce such effects when the instructions are followed. According to the producer, the employees did not use the chemical as prescribed in the instructions: The employees mixed the chemical in large batches; they did not wear protective suits at all times; and, on occasion, they failed to use masks that filtered the fumes.

In addition to being provided with this information, the judges and attorneys were told the amount the plaintiffs were seeking and the approximate amount of time the trial would consume. They also were informed as to whether the case would be tried by a judge (bench-tried) or by a jury. Finally, they were told whether the trial judge was to be the judge currently hearing the case (out of court) or another judge.

After reading the case the judges were asked to indicate (a) how strongly they would mediate the case and (b) the specific techniques they would use. At the end of the questionnaire they were asked about (c) their attitude toward judicial involvement in settlement negotiations and (d) the state in which they practiced.

Analagous questions were posed for the attorneys. They were asked, first, (a) how strongly the judge should mediate (facilitate) settlement of the case and (b) which techniques he or she should use, and, at the end of the questionnaire, about their (c) general attitude toward judicial involvement in settlement negotiations and (d) the state in which they practiced.

Independent Variables

In the sample case, the judge or attorney was informed that (a) the plaintiffs were seeking (case size) \$30,000, \$500,000, or \$30,000,000, (b) the trial time was 5, 15, or 25 days, (c) the trier of the case was a judge (bench-tried) or jury (jury-tried), and (d) the trial judge was to be the current one or a different judge.

In order to glean the judges' and attorneys' attitudes toward judicial involvement in settlement negotiations, they were asked at the end of the questionnaire to answer the question, "In general, how strongly should judges facilitate the settlement of cases?" on a 7-point bipolar scale (1 = *not at all*; 7 = *very strongly*).

In order to test hypothesis 5 (see Hypotheses section), the responses from the judges and attorneys in the 50 states were grouped into 4 sections: Northeast, West, Northcentral, and South (Bureau of Census, 1983).

Dependent Variables

The strength of the judges' mediation efforts (and the attorneys' preference for that involvement) was measured along five facets.

First, after reading the case the judges were asked to answer the question, "In this case, how strongly would you facilitate settlement?" on a 7-point bipolar scale (1 = *not at all*; 7 = *very strongly*). Attorneys were asked an analogous question, "In this case, how strong would you want the judge's settlement efforts to be?" Their answer, "I would want the judge to facilitate settlement _____" was also rated on a 7-point scale (1 = *not at all*; 7 = *very strongly*).

Next, the judges and attorneys were presented with 20 mediation techniques that could be used in the case. Judges checked the techniques they would use, and attorneys checked the ones they would want a judge to use. The number of techniques checked served as the second dependent variable.

The techniques available to the judges and attorneys had previously been selected via a factor analysis (a principal-components analysis with varimax rotations) from the 71 mediation techniques that judges were found to use in two previous studies. These studies of actual settlement negotiations (Wall & Rude, 1985) reveal that judges use many of the 71

Table 1
Judicial Mediation Techniques

Technique
Aggressive
Pressure the ill-prepared attorney
Require a settlement conference even though one is not mandated by court rules
Downgrade the merits of the stronger case and/or the demerits of the weaker
Talk to each lawyer separately about settlement
Set an inexorable trial date
Client-Oriented
Suggest a settlement figure to the client
Note, for the client, the rewards of pretrial settlement
Point out, to the client, the strengths and weaknesses of his case
Speak personally with the client to persuade him or her to accept
Bring the client to the conference
Logical
Offer alternative proposal not thought of by lawyers
Argue logically for concessions
Analyze the case for a lawyer
Offer advice to a lawyer
Suggest a settlement figure after asking for lawyers' inputs
Paternalistic
Ask both lawyers to compromise
Call a certain figure reasonable
Channel discussions to areas that have the highest probability of settlement
Note to a lawyer the high risk of going to trial
Inform the attorneys about the way in which similar cases have been settled

techniques together in four strategic combinations: aggressive, client-oriented, logical, and paternalistic. For the current study, five techniques, which loaded strongly on each factor, were chosen from each grouping and proffered in a random order to the judges and attorneys (see Table 1).

The number of aggressive and client-oriented techniques (i.e., the stronger techniques) chosen by the respondents were considered as the third and fourth dependent variables.

The final dependent measure was the assertiveness of the techniques selected by the judges and attorneys. In order to develop this measure, 300 state judges (selected randomly from a nationwide listing by the National Judicial College) were asked to score each technique on a 7-point bipolar scale (1 = *not at all assertive*; 7 = *extremely assertive*). Their responses (53% response rate) provided an assertiveness rating for each technique. In order to calculate the assertiveness of the subjects in this study, we averaged the assertiveness ratings for the techniques that each judge or attorney used and labeled this dependent measure *technique assertiveness*.

Design and Analysis

A multivariate analysis of variance (MANOVA) with three levels of case size (\$30,000, \$500,000, and \$30,000,000), three levels of trial time (5, 15, or 25 days), two levels of trier of case (judge or jury), two levels of judges (current judge or another judge), four sections of the country (Northeast, West, Northcentral, and South), two sets of respondents (judges and attorneys), and the respondents' attitudes toward judges' involvement in settlement was used to study the judicial mediation.

As noted earlier, the dependent variables included the judges' and attorneys' responses as to (a) how strongly they would (or wished for the judge to) facilitate settlement, (b) the overall number of techniques to be used, (c) the number of aggressive and client-oriented techniques selected, and (d) average assertiveness of the techniques employed.

Results

Of the 650 judges surveyed, 379 responded, for a response rate of 58%; the rate for the attorneys was 73%, with 799 out of 1,100 responding.

The analysis of the judges' and attorneys' responses to the cases reveals that three factors—judges' and attorneys' attitudes toward judicial involvement in settlement, region of the country, and respondent (judge or attorney)—strongly affected the dependent variables.

The respondents' attitudes as predicted in Hypothesis 4 (see Hypotheses section), strongly affected, multivariate $F(5, 998) = 261.14$, $p < .001$, the judges' mediatory efforts and the attorneys' preferences for them. As the respondents' preference for judicial intervention increased, so did (a) the involvement in the case, (b) the number of techniques used, (c) the number of aggressive techniques, (d) the number of client-oriented techniques, and (e) the degree of technique assertiveness. Table 2 presents the correlations between the respondents' attitude and these dependent variables as well as the correlations among the dependent variables.

With regard to sectional differences, it can be seen in Table 3 that respondents from the northeastern and western states chose (a) stronger involvement in the case, (b) more techniques, and (c) more aggressive techniques. Western respondents chose more client-oriented techniques than did respondents from other regions. Northeastern, western, and northcentral respondents chose techniques that were more assertive than did their counterparts in the southern states.

These results support Hypothesis 5, multivariate $F(15, 2755) = 8.78$, $p < .001$; likewise, the analysis of the judges' versus attorneys' data (Table 3) supports Hypothesis 6. The means

Table 2
Correlations Among Respondents' Attitude Toward Judicial Involvement and Dependent Variables

Variable	1	2	3	4	5	6
1. Respondents' attitudes	—	.74	.51	.37	.27	.28
2. Strength of involvement		—	.51	.36	.25	.34
3. Number of techniques used			—	.63	.64	.33
4. Number of aggressive techniques				—	.25	.22
5. Number of client-oriented techniques					—	.10
6. Technique assertiveness						—

Note. Respondents' attitude is an independent variable; all others are dependent. All correlations are significant at $p < .001$.

Table 3
Sectional and Respondent (Judges vs. Attorneys) Effects

Measure	Variable				
	Strength of involvement	Number of techniques used	Number of aggressive techniques	Number of client-oriented techniques	Technique assertiveness
Section					
Northeast	5.2	7.1	2.0	0.68	3.5
West	5.2	7.9	1.7	1.58	3.5
Northcentral	4.6	6.1	1.5	0.76	3.5
South	4.6	5.4	1.2	0.64	3.4
Respondents					
Judges	4.7	6.3	1.4	0.68	3.4
Attorneys	4.9	6.6	1.6	0.96	3.5

reveal that, compared with judges, attorneys prefer stronger facilitation by judges, the use of more techniques, the use of aggressive and client-oriented techniques, and the use of more assertive techniques, multivariate $F(5, 998) = 6.87, p < .001$.

For the elements of the case—size, length, trier, and judge—only the size had a significant effect across all the dependent variables, multivariate $F(10, 1996) = 2.35, p < .01$; the strength of the judges' mediation efforts (and the attorneys' preferences) increased as the size of the case decreased (see Hypothesis 1).

Rather than having a strong main effect, as predicted in Hypothesis 2, trial length interacts with case size to affect judicial involvement. This interaction (see Table 4) indicates that judges base their mediation efforts (and that attorneys base their preferences) somewhat on the trial length of a case, but more on the consistency between the trial length and the case size, $F(4, 1113) = 2.53, p < .04, \omega^2 < .01$.

For a \$30,000,000 case, 15 or 25 days of trial time seems to be deemed acceptable; consequently, the judges do not mediate strongly (nor do the attorneys wish for them to do so). Yet for the smaller cases (\$30,000 and \$500,000) these times seem excessive, and judicial involvement increases.

Hypothesis 3 was not supported by the data; rather, judges who knew they were to decide the case subsequent to their mediation were just as aggressive in mediating it as (a) those who were not scheduled to rule on it or (b) those who would oversee a jury trial. We consider this finding odd, given that several states have canons that prohibit judges' involvement in cases that they are to bench try. We also found that many judges and

attorneys remarked on their questionnaire that judicial involvement in a case scheduled for subsequent bench trial before the same judge was inappropriate.

Offsetting these orientations, perhaps, was the conclusion by some judges that they should initiate the resolution of the case out of court if they were eventually going to decide it in the courtroom.

Discussion

In sum, five variables under study—attitudes toward judicial involvement, region of the country, respondent, size of case, and trial length—affected judicial involvement, and two—trial judge and trier of case—did not. These findings enhance our understanding of the judicial mediation process and contribute to a rudimentary base for guiding judges' efforts.

Over the past six years we have found that many judges are actively mediating the out-of-court settlement of civil cases because they believe that judicial involvement facilitates agreement. Data from the current study allow us to build on our earlier findings; now we can additionally hypothesize that personal and situational variables affect judicial mediation and that judicial mediation, in turn, leads to out-of-court settlement. Further expansion of this model requires us to proceed along three avenues. First, because attitudes have a significant effect upon judges' mediations, it seems worthwhile to ferret out the determinants of these attitudes. Second, investigations should be undertaken to chart the effects that additional variables have on judges' involvement in settlement and the tactics that they employ.

Finally, we need to examine the effects of judges' endeavors on the quality as well as on the speed and frequency of settlement. Within-subject designs seem most appropriate for these investigations, because interjudge variance in attitudes (which has strong effects on their involvement in settlement) is quite high. Prior to its assignment to a judge, each case could be designated randomly as either (a) one in which the judge is to facilitate settlement or (b) one in which the judge is to forgo settlement efforts. For cases with the former designation, the judge would indicate the techniques he or she used. For both the "facilitate settlement" and "forgo settlement" cases, researchers could record the case characteristics (size, probable trial length,

Table 4
Trial Length by Case Size Interaction On Judges' Involvement

Trial length (in days)	Case size		
	\$30,000	\$500,000	\$30,000,000
5	4.9	4.4	4.8
15	5.0	4.9	4.3
25	5.2	5.1	4.7

Note. Strength of judges' involvement measured on bipolar scale (1 = not at all; 7 = very strong).

number of parties to the dispute, and so forth), whether or not the case was settled, and the speed of settlement. Subsequent interviews with the adversaries could provide measures of their satisfaction with the settlement process and outcomes.

These studies, in combination with earlier ones, will provide a basis for educating judges about the conditions under which their colleagues facilitate settlement. They also will guide judges in their own mediations by pointing out (a) which techniques are most apt to reap settlement and (b) the conditions under which each technique is most likely to prove successful.

References

- Boyum, A. (1979). A perspective on civil delay in trial courts. *Justice System Journal*, 5, 170-185.
- Brazil, W. D. (1985). *Settling civil disputes*. Chicago: American Bar Association.
- Bureau of the Census. (1983). *1980 census of population and housing: Geographic identification code scheme*. Washington, DC: U.S. Government Printing Office.
- Carter, J. M. (1962). Effective calendar control: Objectives and methods. *Federal Rules Decisions*, 29, 227-240.
- Flanders, S. (1977). *Case management and court management in the United States district courts*. Washington, DC: Federal Judicial Center.
- Galanter, M. (1985). A settlement judge, not a trial judge: Judicial mediation in the United States. *Journal of Law and Society*, 12, 1-18.
- Haynes, J. M. (1981). *Divorce mediation: A practical guide for therapists and counselors*. New York: Springer.
- Kochan, T. A., & Jick, T. (1978). The public sector mediation process: A theory and empirical examination. *Journal of Conflict Resolution*, 22, 209-240.
- Kolb, D. M. (1983). *The mediators*. Cambridge, MA: MIT Press.
- Schiller, L. F., & Wall, J. A. (1981). Judicial settlement techniques. *The American Journal of Trial Advocacy*, 5, 39-61.
- Title, C. (1981). The lawyer's role in settlement conferences. *The American Bar Association Journal*, 67, 592-597.
- Touval, S. (1982). *The peace brokers: Mediators in the Arab-Israeli conflict 1948-1979*. Princeton, NJ: Princeton University Press.
- Wall, J. A. (1981). Mediation: An analysis, review, and proposed research. *Journal of Conflict Resolution*, 25, 157-180.
- Wall, J. A., & Rude, D. E. (1985). Judicial mediation: Techniques, strategies, and situational effects. *Journal of Social Issues*, 41, 47-63.
- Wall, J. A., & Schiller, L. F. (1982). Judicial involvement in pretrial settlement: A judge is not a bump on a log. *The American Journal of Trial Advocacy*, 6, 27-45.
- Wall, J. A., Schiller, L. F., & Ebert, R. J. (1984). Should judges grease the slow wheels of justice? A survey on the effectiveness of judicial mediary techniques. *The Journal of American Trial Advocacy*, 8, 83-114.
- Watson, G. D. (1984, April). *Judicial mediation: The results of a controlled experiment in the use of settlement-oriented pretrial conferences*. Paper presented at the 1984 Law and Society Meeting, Boston.
- Young, O. R. (1972). Intermediaries: Additional thoughts on third parties. *Journal of Conflict Resolution*, 16, 51-65.

Received May 5, 1986

Revision received June 30, 1986 ■

Cognitive Categorization and Quality of Performance Ratings

Michael K. Mount and Duane E. Thompson

Department of Industrial Relations and Human Resources, University of Iowa

The effects of cognitive categorization of raters on accuracy, leniency, and halo of performance evaluations were investigated in a field setting. One hundred seventy-four subordinates evaluated the performance of their managers on three performance dimensions. Managers were categorized as congruent or incongruent based on subordinates' perceptions of the extent to which the manager's behavior met the subordinate's expectations. The results indicated that the quality of ratings assigned by subordinates was related to the cognitive categories used. As hypothesized, ratings of managers who were categorized as congruent were found to be more accurate and also to contain more leniency and halo tendency than the ratings of managers who were categorized as incongruent. Implications of these findings for performance-appraisal research are discussed.

Researchers have long been interested in identifying methods of improving the quality and accuracy of performance ratings. One popular approach has been to manipulate the content and format of the rating system. In general, the results of these studies have been unsuccessful and have led several researchers to call for a moratorium on such research and a redirection of efforts toward the study of the cognitive processes involved (Cooper, 1981; Feldman, 1981; Ilgen & Feldman, 1983; Landy & Farr, 1980).

The concept that plays a key role in understanding the cognitive processes of raters is categorization (Feldman, 1981; Ilgen & Feldman, 1983). Briefly stated, the central concept is that individuals perceive and process information in terms of abstract categories defined by various schemata and, in some cases, by somewhat more concrete prototypes. These categorization schemata or prototypes, which may be based on formal or informal sources of information, allow individuals to achieve "cognitive economy" by reducing the amount of information processed and stored (e.g., Behling, Gifford, & Tolliver, 1980; Smith, Adams, & Schorr, 1978). Categorization of individuals may proceed under one of two conditions. When the observed behavior of an individual is congruent with expectations, as defined by the categorization schema or prototype, it is noted and categorized automatically—a consistent mapping condition. But when the observed behavior is inconsistent with categorization schemata or prototypical expectations, conscious attention must be used to categorize the individual's behavior—a variable mapping condition (Ilgen & Feldman, 1983). Categorization schemata and prototypes play a key role in either condition. The actual behavior of an individual is compared with the categorization schema or prototype, and, once categorized, further

perception and recall of the individual's behavior is biased toward the category (Cantor & Mischell, 1977).

The effects of the cognitive processes associated with categorization on halo, leniency, and accuracy are of particular interest. For example, according to Feldman (1981) halo is the result of a heuristic process in which information is stored automatically as part of a prototype-based category. Halo in performance ratings is produced because distinct but similar behavior patterns are treated as being equivalent when classified into the same category (Nathan & Alexander, 1985; Nathan & Lord, 1983). Thus, halo is expected to occur when the observed behaviors of a ratee are consistent with the categorization schema of the rater. This is especially true with trait ratings because traits are recalled as a "bundle" and covary with the category (Feldman, 1981).

Furthermore, leniency and stringency may be influenced by the categorization schema because the category to which an individual is assigned may have a positive or a negative connotation associated with it, which in turn influences the way ratee behaviors are recalled. When the categorization schema is defined in terms of desired behaviors, it is logical to assume that the category takes on a positive connotation. Conversely, when a categorization schema is defined by behaviors that are incongruent with the rater's desires the category takes on a negative connotation. If a ratee is perceived as fitting a positive categorization schema, leniency is likely to result. If an individual is perceived as inconsistent with the positive categorization schema or as fitting a negative categorization schema, stringency would be expected.

Feldman (1981) also suggests that the same processes that produce halo and leniency can account for accuracy. There is some empirical evidence, for example, that category-consistent information is more likely to be recalled than category-inconsistent information (Hastie & Kumer, 1979; Lingle & Ostrom, 1979). Furthermore, it has been shown that when the rater expects certain behaviors from a stimulus person, those behaviors are noticed and recalled more than unexpected but equally available behaviors (Zadny & Gerard, 1974).

From a conceptual point of view, it is known that raters differ in their understanding of the rater's roles and responsibilities.

Portions of this article were presented at the First Annual Convention of the Society for Industrial/Organizational Psychology, Inc. (Division 14 of the American Psychological Association), Chicago, April 10–11, 1986.

Correspondence concerning this article should be addressed to Michael K. Mount, Department of Industrial Relations and Human Resources, University of Iowa, Iowa City, Iowa 52242.

That is, some raters have better, more differentiated category systems that correspond more closely to the role prescriptions of the ratee. Behaviors exhibited by the rater are observed and stored by the rater. Over time, however, only the category in which the ratee is stored is recalled, rather than the specific behaviors. If the observed behaviors of the ratee are consistent with the features of the category, the behaviors of the ratee will be recalled more accurately because the features of the prototype are also true of the person being evaluated. These findings are consistent with those reported by Zadny & Gerard (1974). In summary, for both empirical and conceptual reasons, we expect a positive relation between category congruence and rating accuracy.

The purpose of this study is to investigate the relation of the cognitive categorization process of raters and the quality of performance ratings in a work setting. The effect that categorizing an individual as consistent or inconsistent with a desired schema has on the accuracy, halo, and leniency or stringency of performance evaluations is specifically investigated. Three hypotheses are investigated.

1. Ratings of managers who are perceived as congruent with a desired categorization schema will be more accurate than ratings of those who are not so perceived.
2. Ratings of managers who are perceived as congruent with a desired categorization schema will contain more halo than ratings of those who are not so perceived.
3. Ratings of managers who are perceived as congruent with a desired categorization schema will be more lenient than ratings of those who are not so perceived.

Additionally, the study investigates the relations among three situation-specific variables and the accuracy, leniency, and halo present in the ratings. Bernardin and Beatty (1984) have suggested that it may be fruitful to explore variables such as the rater's experience, the extent to which the ratee's behavior can be observed, and the knowledge the rater possesses about the ratee's performance. Three situation-specific variables are investigated in this study: (a) the length of time the subordinate had worked for the manager, (b) how often the manager discussed performance with the subordinate, and (c) the length of time the manager had been in a management position within or outside the organization. No specific hypotheses are offered for these variables.

Method

Sample

The initial sample consisted of 255 midlevel managers (hereafter referred to as *managers*), their supervisors ($n = 255$), and their subordinates ($n = 918$). The managers were randomly selected from the population of managers residing in the same metropolitan area of the organization's corporate headquarters. The managers were selected from the first and second levels of management only in order to minimize sources of unreliability due to different organizational levels (Borman, 1974; Pulakos & Wexley, 1983). The final sample consisted of only those managers who had at least 4 subordinates reporting to them, each of whom had worked for the manager for at least 1 year. (The requirement of 4 subordinates per manager was necessary in order to obtain a measure of the true performance level for each manager as explained in the *Rating Measures* section below.) These restrictions reduced the final sample to

174 managers. The managers had been in their present jobs an average of 2.5 years, had been managers an average of 6.8 years, and had worked for their current supervisor an average of 2.4 years.

Procedure

The data were collected as part of a management development study in the organization. Although participants were asked to indicate their employee identification numbers they were assured the results would be strictly confidential. The subjects were introduced to the research through a letter from the vice-president for public affairs and personnel, who requested the participation of the managers, their subordinates, and their supervisors and explained that the study would be administered by an external consulting firm to ensure confidentiality. Participants were instructed to complete and return the questionnaires to the consultant within 10 days. The return rate was 86% for supervisors, 88% for subordinates, and 90% for managers.

Instruments

A modified version of the Leadership Analysis Questionnaire (LAQ; e.g., Davis & Mount, 1984a; Mount, 1984a, 1984b) was used to assess managerial effectiveness. There were three versions of the questionnaire: one for the managers' self evaluation, one for the immediate supervisors of the managers, and one for the subordinates of the managers.

Background information. The first section of the questionnaire consisted of items designed to assess background information about the respondent. Items included the individual's company-employee identification number, the length of time the respondent had worked in the current job, the length of time the individual had supervised or worked for the manager, and so forth.

Behavioral performance items. The second section contained 14 behavioral items designed to measure managerial job performance. These statements were based on a study by Tornow and Pinto (1976) that developed a Management Position Description Questionnaire (MPDQ) for objectively describing the content of management positions in behavioral terms. (For a complete description of the development of the MPDQ, see Tornow and Pinto.)

Supervisors and subordinates were asked to rate the performance effectiveness of the target managers on each of the 14 behavioral statements using a 9-point scale ranging from low (1–3) to medium (4–6) to high (7–9). Managers were asked to provide self ratings on the same behavioral statements. Thus, self ratings, supervisor ratings, and subordinate ratings were obtained for the 174 managers on each of the 14 behavioral statements. Each item was assigned to one of three performance dimensions identified through factor analysis.

Perceived role congruence. The subordinates' version of the LAQ contained a third section designed to measure manager-role congruence as perceived by subordinates. That is, it measured the extent to which the subordinate perceived the manager to be performing responsibilities in the way the subordinate believed they should be performed. The scale consisted of 10 behavioral items from the MPDQ: (a) providing instructions and explanations; (b) assuring the proper orientation and training of employees; (c) scheduling and assigning work efficiently; (d) encouraging high standards of quality and quantity; (e) providing praise and recognition; (f) correcting employees promptly; (g) conducting performance appraisals regularly; (h) exercising tact and sensitivity when dealing with others; (i) showing genuine concern for employees' job satisfaction; and (j) working to reduce interpersonal conflicts fairly. For each of these statements the subordinate was instructed to indicate the amount of the activity the manager *should* display compared with the amount *currently* displayed. A 5-point scale was used (1 = *less than currently displayed*, 3 = *same as currently displayed*, and 5 = *more than currently displayed*). The coefficient alpha reliability estimate for this scale was .88.

Classification Variable

Subordinates' perceived congruence (SPC) was the classification variable and consisted of two levels: high congruence and low congruence. SPC was computed using the Index of Profile Similarity, D (Cronbach & Gleser, 1953; Osgood, Suci, & Tannenbaum, 1957). The responses of one randomly selected subordinate to the 10 behavioral items described in the preceding paragraph were used to compute subordinate's perceived congruence (SPC). For these computations, the 5-point response scale for these items was scored as follows: $-2 = \text{less than currently displayed}$, $0 = \text{the same as currently displayed}$, and $2 = \text{more than currently displayed}$. The SPC value was then calculated as the square root of the sum of squared scores. Thus, a low SPC score indicated that the manager was performing in a way consistent with the subordinate's expectations, and a high score indicated that the manager was performing in a way inconsistent with the subordinate's expectations, that is, doing too much or too little on the performance dimensions. The possible values of SPC ranged from 0 to 6.32 ($\sqrt{40}$). Managers whose SPC scores fell below the median ($M = 2.30$) were defined as high congruence and those above the median were defined as low congruence.

Rating Measures

Three measures of the psychometric quality of subordinate ratings (accuracy, leniency, and halo) were investigated for each of the three performance dimensions. Investigations of rater accuracy and leniency require measures of true scores. In laboratory or field settings one may only approximate true scores. In this study, "true performance levels" were calculated for each manager in the following way. First, the randomly selected subordinate who provided the SPC score for a manager was removed from this step of the analysis. Next, the average of the remaining subordinate ratings was computed for each item for each manager. Then, the three rating constituencies (supervisor, self, and the average subordinate ratings) were unit weighted, summed, and divided by three to obtain the true performance level for each item.

One may assume that each rating constituency contains some portion of variance that is common and some portion of variance unique to the other sources. A number of studies have shown that the average of several ratings is more reliable than a single rating (e.g., French & Bell, 1978; Latham, Fay, & Saari, 1979; Miner, 1968). The true performance measure used in the study is not the true score of the individual, but is a better estimate of the individual's performance than a single rating source alone. This procedure is similar to that employed in laboratory research in which expert ratings were averaged to obtain "true scores" (Bernardin, 1978; Borman, 1979b). Given the field setting of this research, the calculation of true performance levels is believed to be a strength of the study.

Accuracy was defined as the average absolute difference between the randomly selected subordinate's evaluation of the manager's performance and the manager's true performance level for the items making up a performance dimension. This definition is similar to that used by other researchers (e.g., Bernardin & Pence, 1980; Heneman & Wexley, 1983).

Leniency was defined as the average difference between the randomly selected subordinate's evaluation of the manager's performance and the manager's true performance level on the behavioral items making up a performance dimension.

Halo was defined as the standard deviation of the randomly selected subordinate's ratings of the behavioral items making up a performance dimension.

Thus, accuracy, leniency, and halo measures were obtained on each of the three performance dimensions for one randomly selected subordinate for each manager.

The three situation-specific variables investigated in this study were defined using three criteria: (a) the length of time the subordinate had

Table 1

Varimax Rotated Factor Pattern Matrix on Managerial Performance Dimensions

Managerial effectiveness item	General supervision	Motivation	Expertise
Processing paperwork	.79	.22	-.03
Developing new solutions	.68	.17	.17
Applying innovative procedures	.64	.09	.30
Maintaining managerial control	.57	.39	.25
Recommending training programs	.53	.01	.38
Documenting decision	.50	.29	.22
Keeping informed of latest developments	.15	.84	.04
Giving stimulating assignments	.05	.81	.11
Monitoring employee progress	.35	.68	.26
Capitalizing on job changes	.29	.53	.43
Setting an example	.37	.50	.34
Answering difficult questions	.13	.20	.76
Recording important details	.15	.24	.75
Serving as a resource person	.35	.63	.54

worked for the manager (1 = *less than 1 year* to 6 = *11 or more years*), (b) how often the manager discussed performance with the subordinate (1 = *once a week* and 4 = *once a year*), and (c) the length of time the manager had been in a management position within or outside the organization (1 = *less than 1 year* to 6 = *16-20 years*).

Data Analysis

The data were analyzed using multivariate analyses of covariance (MANCOVAs). Level of congruence (high or low) was the classification variable, and the degree of accuracy, halo, and leniency for each of the three performance dimensions was the rating variable. The covariates consisted of the three situation-specific variables. Thus, three MANCOVAs were conducted, one for each of the three rating variables.

Results

Factor Analysis

Table 1 presents the results of the factor analysis of the matrix of intercorrelations (14×14) of the managerial effectiveness items. This factor analysis was performed on mean ratings across raters for each manager on each of the 14 behavioral dimensions. First, the subordinate ratings for each item were averaged for each manager. Then the three rating constituencies (supervisor, self, and subordinate) were unit weighted, summed, and divided by three. The resulting values represent the data that were factor analyzed in the study. The principal-components method was used with unities in the diagonal, and factoring was stopped when eigenvalues were less than unity. Three factors with eigenvalues greater than 1.0 (General Supervision, Motivation, and Expertise) were extracted, and were found to account for 56.3% of the common factor variance associated with subordinates' responses to the questionnaire. Respective coefficient alpha reliabilities ranged from .78 to .84.

Multivariate Analysis of Covariance

The means and standard deviations for accuracy, halo, and leniency for the three performance dimensions are shown in

Table 2
Means and Standard Deviations for Rating Measures By Congruence Level

Congruence level	General supervision		Motivation		Expertise	
	M	SD	M	SD	M	SD
Accuracy						
Congruent	4.04	1.01	2.04	1.04	3.16	1.04
Incongruent	4.36	1.09	2.40	1.07	3.47	1.15
Halo						
Congruent	1.43	0.52	1.30	0.83	1.27	0.45
Incongruent	1.62	0.59	1.55	0.75	1.41	0.53
Leniency						
Congruent	0.34	0.67	0.39	0.64	0.37	0.60
Incongruent	0.11	0.65	0.15	0.73	0.14	0.64

Note. N = 87.

Table 2. The relation between level of congruence and rating accuracy, and between halo and leniency are shown in Table 3. The results of the MANCOVA indicated a statistically significant effect, $F(3, 172) = 2.68, p > .05$, using Wilks's Criterion. None of the covariates was found to be statistically significant, and, consequently, only results for main effects will be reported. Inspection of the univariate analyses of variance (ANOVAS) indicates that Hypothesis 1 (ratings of managers who are perceived as congruent with a desired categorization schema will be more accurate than ratings of those who are not so perceived) was confirmed for each of the three performance dimensions ($p < .01$ in all cases, and ω^2 ranged from .03 to .06). Rating accuracy was higher when the subordinate perceived the manager to be performing in a way congruent with the subordinates' expectations.

The results of the MANCOVA for halo tendency also indicated a significant effect, $F(3, 172) = 2.82, p < .03$, using Wilks's Cri-

terion. As before, none of the three covariates were found to be significant, and, consequently, only results for main effects are reported. Inspection of the univariate ANOVAS indicates consistent effects across the three performance dimensions ($p < .05$, and $\omega^2 = .03$, in all cases), as shown in Table 3. The results are consistent with Hypothesis 2: Subordinates who perceive their managers to be congruent exhibit greater halo on the three performance dimensions than subordinates who perceive their managers to be incongruent.

The MANCOVA assessing leniency effects also indicated a highly significant effect, $F(3, 172) = 6.87, p < .0001$, using Wilks's Criterion. As before, none of the three covariates were found to be significant. The effects for the univariate ANOVAS are consistent across each of the three performance dimensions ($p < .01$ in all cases and ω^2 ranged from .04 to .06) as shown in Table 3. It can be seen that there is a significant effect in the expected direction for level of congruence. These data support

Table 3
Analyses of Variance for Rating Measures by Congruence Level

Source	General Supervision				Motivation				Expertise			
	df	MS	F	ω^2	df	MS	F	ω^2	df	MS	F	ω^2
Rating measure: Accuracy												
Congruence	1	353.02	7.67**	.04	1	70.54	11.62***	.06	1	131.39	8.52**	.03
Error	172	46.00			172	6.07			172	15.42		
Rating measure: Halo												
Congruence	1	53.45	6.35**	.03	1	16.52	5.66*	.03	1	32.47	5.74*	.03
Error	172	8.42			172	2.92			172	5.66		
Rating measure: Leniency												
Congruence	1	611.64	11.54***	.06	1	165.57	7.26**	.04	1	416.39	10.39***	.05
Error	172	52.99			172	22.80			172	40.09		

* $p < .05$. ** $p < .01$. *** $p < .001$.

Hypothesis 3: Subordinates who perceive their managers to be congruent exhibit greater leniency on the three performance dimensions than subordinates who perceive their managers to be incongruent.

Discussion

The central issue investigated in this study was the way the cognitive categorization process is related to the quality of performance ratings. The results indicate that the nature of the categorization is clearly related to the degree of accuracy, halo, and leniency in the ratings.

First, a note is in order regarding the congruence measure used in the study. A question that must be addressed is whether the congruence measure actually measures something different than the performance of the manager. In order to address this issue, correlations were obtained between the measure of congruence (SPC) for the manager (congruent or incongruent) and the subordinate's ratings of performance on each of the three performance dimensions. The biserial correlations were found to be nonsignificant for each dimension. Thus, the congruence measure appears to measure something conceptually different from performance.¹

Perhaps the most interesting results in the study relate to rating accuracy. The findings indicate that the ratings are more accurate when the behaviors of the ratee are consistent with the expectations of the rater. One possible explanation for this is that behaviors that are expected are more salient and, as a result, are noticed and recalled more easily than unexpected behaviors (Zadny & Gerard, 1974). The more prototypical the behavior of the ratee, the greater the number of salient cues on which the subsequent evaluation is based.

Another potential explanation for greater accuracy when behaviors of the ratee are consistent with the expectations of the rater is that the reliability of the performance ratings was higher among raters in the high congruence group compared with the low congruence group. In order to test this, the standard deviation of the performance scores across the supervisor, self, and subordinate ratings was computed for each manager on each of the three performance dimensions for both congruent and incongruent groups. (Greater standard deviations mean less agreement or reliability among rating sources.) The results of three ANOVAs in which SPC was the classification variable and the standard deviation on the three performance dimensions for each manager was the dependent variable, indicated no significant differences between the groups on any of the three performance dimensions. These findings indicate no difference in the degree of reliability (agreement) on the performance ratings between congruent and incongruent groups. Thus, the greater accuracy observed for the congruent group does not appear to be explained by differences in reliability of the performance measures.

The results of the study also indicate that when the behaviors of the ratee are consistent with the expectations of the rater, the ratings are more lenient and contain greater halo. In general, the findings regarding halo are consistent with those reported in other studies that indicate that once a ratee is categorized further perception and recall of the ratee is biased toward the category (Cantor & Mischel, 1977; Snyder & Swann, 1978).

Ratees are recalled as part of a category prototype that represents the central tendencies of the category. The more prototypical the ratee, the greater the likelihood of halo bias. Ratees whose behavior is consistent with rater expectations will be recalled with greater halo because the behavior is more prototypical than ratees whose behavior is inconsistent.

Leniency bias may be explained in an analogous way. Recall of ratee behavior may be influenced positively or negatively by the qualities connoted by the category. When the behavior of the ratee is perceived to be consistent with rater expectations the category takes on a positive connotation because the behaviors are viewed as acceptable or expected. On the other hand, when the behavior of the ratee is perceived as incongruent the category has a negative connotation because the behaviors are unacceptable or unexpected. Subsequent recall and evaluation of ratee behaviors will be done in a way consistent with the category, and behaviors that have been categorized as positive will be evaluated with greater leniency than those categorized as negative. This explanation is admittedly speculative, however, and alternative explanations may certainly be appropriate.

These findings have a number of important implications for performance-appraisal practices. They suggest that efforts to improve rating accuracy should be directed toward identifying factors that influence the types of categories used by raters. There is evidence, for example, that expertise tends to create more differentiated category systems (Ilgen & Feldman, 1983; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Moreover, several researchers have suggested that training may be helpful in teaching raters a common nomenclature or common frame of reference for defining performance effectiveness levels (Bernardin, 1979, 1981; Borman, 1979a). In frame-of-reference training, raters who have been identified as possessing idiosyncratic work standards are presented with normative standards to compare with their own (Bernardin, 1979, 1981; see also Davis & Mount, 1984b.) Group problem-solving techniques and group discussions are used to develop standard schemata among raters. Several initial tests indicate that this may be a promising method for increasing rater accuracy (Bernardin, Cardy, & Abbott, 1982). One explanation for the success of the method is that a standard performance schemata is imposed on the raters before the observation period, which results in the development of similar categories among raters. It has been shown, for example, that increasing rater's knowledge of the dimensions prior to the rating task increases rating accuracy (Bernardin et al., 1982).

The results discussed above do not directly assess the relations among the measures of accuracy, halo, and leniency for individual raters. In order to investigate this the three rating measures were correlated for the congruent and incongruent groups of subordinates. The results indicate a consistently positive correlation between the degree of accuracy and halo among the three performance dimensions for both congruent and incongruent groups. No consistent relations were found between the measures of leniency and accuracy or between leniency and halo on the three performance dimensions. These results seem

¹ We would like to acknowledge the comments of two anonymous reviewers regarding this point.

paradoxical on the surface because ratings that are more accurate also contain more halo. Halo was defined as the standard deviation of the ratings on items contained within the three dimensions identified through factor analysis. It would be expected that ratings of items within a factor would be moderately to highly correlated, which would result in a relatively high degree of halo within each factor. However, this does not explain the high correlation between the degree of halo and accuracy in the ratings.

The traditional belief among industrial/organizational psychologists is that halo represents a source of error that reduces the degree of accuracy in performance ratings. As several researchers have indicated recently, however, the absence of rating errors such as central tendency, leniency, halo, and so forth, does not necessarily imply more accurate ratings (Bernardin & Pence, 1980; Borman, 1978; Ilgen & Feldman, 1983). In fact, the absence of such tendencies may not be possible given the cognitive limitations of appraisers (Bernardin & Beatty, 1984; Ilgen & Feldman, 1983). The obvious implication of these findings is that rating accuracy should be assessed independently of other measures such as halo.

No evidence was obtained in the study regarding the role of the situation-specific variables for any of the rating measures, including accuracy. These findings may be attributable to the restricted sample used in the study, which included only those managers and subordinates who had worked together for at least one year. The results presented here indicate that the length of time the subordinate had worked for the manager, how often the manager and subordinate discuss performance, and the length of time the manager had been in a management position were not related to the rating quality measures. It is not known, however, if the same results would hold true for subordinates who had worked for their manager for less than one year.

In this study the two cognitive categories, congruent and incongruent, were created somewhat arbitrarily. It should be noted, however, that these categories correspond to the dual-process system of evaluation and decision making discussed by Feldman (1981) and Ilgen and Feldman (1983). When the behavior of the ratee is consistent with the expectations or prototype of the rater, it is categorized automatically. Because the automatic process is insufficient to encompass all categorizations, Ilgen and Feldman (1983) suggested that a switching mechanism is necessary to bring the controlled process into play. They postulated a hypothetical mechanism that matches incoming information with the available prototype or schema. As in the case of this study, when the "threshold of discrepancy" is exceeded, the mechanism calls for a controlled categorization. Although direct evidence for this mechanism is minimal (Hastie, 1978; McCloskey & Glucksberg, 1978, 1979; Tversky, 1977), the procedure used in the study is consistent with current theoretical conceptualizations of the categorization process.

Another potential limitation in the study is the way the true performance level was defined. As mentioned earlier, it is rare, if not impossible, to obtain a measure of an individual's true score. Previous research in laboratory settings has used the average rating by "experts" as the measure of the true score. Given the constraints that exist in an organizational setting, however,

it is believed that the method used in this study is a reasonable proxy for the individual's true score.

Finally, it should be noted that although the effects of congruence were statistically significant in the expected direction for all dependent variables, the magnitude of variance accounted for was relatively small (ω^2 ranged from .03 to .06).

References

- Behling, V. W., Gifford, W. E., & Tolliver, J. M. (1980). Effects of grouping information on decision making under risk. *Decision Sciences*, 11, 272-283.
- Bernardin, H. J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. *Journal of Applied Psychology*, 63, 301-308.
- Bernardin, H. J. (1979). Rater training: A critique and reconceptualization. In R. C. Huseman (Ed.), *Proceedings of the 39th Annual Meeting of the Academy of Management*, 216-220.
- Bernardin, H. J. (1981, August). *Rater training strategies: An integrative strategy*. Paper presented at the 89th annual convention of the American Psychological Association, Los Angeles.
- Bernardin, H. J., & Beatty, R. W. (1984). *Performance appraisal: Assessing human behavior at work*. Boston: Kent.
- Bernardin, H. J., Cardy, R. L., & Abbott, J. (1982). *The effects of individual performance schemata, familiarization with the rating scales, and rater motivation on rating effectiveness*. Paper presented at the 43rd annual meeting of the Academy of Management, New York.
- Bernardin, H. J., & Pence, E. C. (1980). Rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology*, 65, 60-66.
- Borman, W. C. (1974). The ratings of individuals in organizations: An alternate approach. *Organization Behavior and Human Performance*, 12, 105-124.
- Borman, W. C. (1978). Exploring upper limits of reliability and validity in performance ratings. *Journal of Applied Psychology*, 63, 135-144.
- Borman, W. C. (1979a). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology*, 64, 410-421.
- Borman, W. C. (1979b). Individual difference correlates of rating accuracy using behavior scales. *Applied Psychological Measurement*, 3, 105-111.
- Cantor, N. G., & Mischel, W. (1977). Traits as prototypes: Effects on recognition memory. *Journal of Personality and Social Psychology*, 35, 38-48.
- Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin*, 90, 218-244.
- Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin*, 50, 456-473.
- Davis, B. L., & Mount, M. K. (1984a). Effectiveness of performance appraisal training with computer assisted instruction and behavior modeling. *Personnel Psychology*, 37, 439-452.
- Davis, B. L., & Mount, M. K. (1984b). Design and use of a performance appraisal feedback system. *Personnel Administrator*, 29(3), 91-97.
- Feldman, J. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology*, 66, 127-148.
- French, W. L., & Bell, C. H., Jr. (1978). *Organization development: Behavioral Science interventions for organization improvement*. Englewood Cliffs, NJ: Prentice-Hall.
- Hastie, R. (1978). Memory for information that is congruent or incongruent with a conceptual schema. In E. T. Higgins (Ed.), *Social cognition* (pp. 95-124). Hillsdale, NJ: Erlbaum.
- Hastie, R., & Kumer, P. A. (1979). Person memory: Personality traits as organizing principles in memory for behaviors. *Journal of Personality and Social Psychology*, 37, 25-38.
- Heneman, R. L., & Wexley, K. N. (1983). The effects of time delay in

- rating and amount of information observed on performance rating accuracy. *Academy of Management Journal*, 26, 677-686.
- Ilgen, D. R., & Feldman, J. M. (1983). Performance appraisal: A process focus. In L. L. Cummings & B. Staw (Eds.), *Research in Organization Behavior* (pp. 141-197). Greenwich, CT: JAI Press.
- Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, 87, 82-107.
- Latham, G. P., Fay, C. H., & Saari, L. M. (1979). The development of behavioral observation scales for appraising the performance of foremen. *Personnel Psychology*, 32, 299-311.
- Lingle, J. H., & Ostrom, T. M. (1979). Retrieval selectivity in memory-based impression judgments. *Journal of Personality and Social Psychology*, 37, 180-194.
- McCloskey, M. E., & Glucksberg, S. (1978). Decision processes in verifying category membership statements: Implications of models of semantic memory. *Cognitive Psychology*, 11, 1-37.
- McCloskey, M. E., & Glucksberg, S. (1979). Natural categories: Well-defined or fuzzy sets? *Memory & Cognition*, 6, 462-472.
- Miner, J. B. (1968). Management appraisal: A capsule review and current references. *Business Horizons*, 11, 83-96.
- Mount, M. K. (1984a). Psychometric properties of subordinate ratings of managerial performance. *Personnel Psychology*, 37, 687-702.
- Mount, M. K. (1984b). Supervisor, self and subordinate rating of performance and satisfaction with supervision. *Journal of Management*, 10, 305-320.
- Nathan, B. R., & Alexander, R. A. (1985). The role of inferential accuracy in performance rating. *Academy of Management Review*, 10, 109-115.
- Nathan, B. R., & Lord, R. G. (1983). Cognitive categorization and dimensional schemata: A process approach to the study of halo in performance ratings. *Journal of Applied Psychology*, 1, 102-114.
- Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.
- Pulakos, E. D., & Wexley, K. N. (1983). The relationship among perceptual similarity, sex, and performance ratings in manager-subordinate dyads. *Academy of Management Journal*, 26, 129-139.
- Rosch, E., Mervis, C. G., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8, 382-439.
- Smith, E. E., Adams, N., & Schorr, D. (1978). Fact retrieval and the paradox of inference. *Cognitive Psychology*, 10, 438-464.
- Snyder, M., & Swann, W. B., Jr. (1978). Hypothesis testing processes in social interaction. *Journal of Personality and Social Psychology*, 36, 1202-1212.
- Tornow, W. W., & Pinto, P. (1976). The development of a managerial job taxonomy: A system for describing, classifying, and evaluating executive positions. *Journal of Applied Psychology*, 61, 410-418.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327-352.
- Zadny, J., & Gerard, H. B. (1974). Attributed attention and information selectivity. *Journal of Experimental and Social Psychology*, 10, 34-52.

Received May 30, 1986

Revision received September 8, 1986 ■

Effects of Raters' Stress on the Dispersion and Favorability of Performance Ratings

Shanthi Srinivas
College of Business Administration
Pennsylvania State University

Stephan J. Motowidlo
Personnel Decisions Research Institute
Minneapolis, Minnesota

We tested effects of raters' stress on the favorability and dispersion of performance ratings. In all, 120 undergraduates completed either a stressful or an unstressful inbasket exercise, either before or after they saw a videotaped portrayal of a manager's job performance. Then they rated the manager on several performance dimensions. Ratings provided by participants who completed the stressful inbasket showed less dispersion across performance dimensions, but no differences on favorability. Whether participants were stressed before or after viewing the performance videotape made no significant difference in the favorability or dispersion of their ratings. This suggests that stressful experiences affect at least the retrieval stage of information processing and might affect the input stage as well. There was a significant interaction with individual differences in Type A pattern. In the unstressful condition, Type A scores were positively correlated with dispersion ($r = .32, p < .05$); in the stressful condition, the correlation was negative ($r = -.15$), but not statistically significant.

Stressful experiences may create a psychological state that endures and affects behavior for some time after the stressor has been removed. S. Cohen (1980) reviewed several studies showing that people subjected to stressful conditions subsequently perform less effectively on tasks that call for tolerance for frustration, clerical accuracy, and the ability to avoid perceptual distractions. They also become less sensitive to others and show "a decrease in helping, a decrease in the recognition of individual differences, and an increase in aggression" (p. 95).

Results like these suggest that stress may also have important effects on performance evaluation. In particular, two of the hypotheses that S. Cohen (1980) advanced to explain effects of stress, the psychic cost hypothesis and the frustration mood hypothesis, describe alternative mechanisms through which stress may influence cognitive processes in performance evaluation.

The psychic cost hypothesis presumes that stressors force people to pay special attention to possible physical or psychological threats. This increased demand on attentional capacity makes it more difficult to perform other activities that also require a great deal of attention. As a result, stressed people may use strategies to simplify the cognitive demands of a task. When the task is to judge performance along several dimensions, one simplifying strategy might be to rely more heavily on the global, evaluative first impression. This suggests that ratings of a single

ratee provided by a stressed rater will show less variability or dispersion across dimensions than those provided by a rater who has not been stressed.

The frustration mood hypothesis suggests that stressors influence behavior through the mediating effects of mood. Stressors are thought to produce frustration, annoyance, and irritation that, in turn, affect interpersonal sensitivity and motivation to perform some kinds of tasks. If a stressor triggers mood states like these, it can also affect information processing by influencing the kind of information an individual attends to and retrieves from memory (Bower, 1981; Isen, Shalke, Clark, & Karp, 1978). This suggests that stressed raters will tend to notice and recall negative information about performers.

The psychic cost hypothesis predicts that in comparison to unstressed raters, stressed raters will produce ratings that are less dispersed across dimensions. The frustration mood hypothesis predicts that stressed raters will rate performance less favorably. The major purpose of this study is to test both predictions.

A second purpose is to determine which stage or stages of information processing are affected by stress. If stress affects performance evaluations, it could do so at the information input stage, the information retrieval stage, or both. This study was designed to explore three possibilities—that stress affects (a) input but not retrieval, (b) retrieval but not input, or (c) both input and retrieval. We hope to rule out one or more of these possibilities.

The third purpose is to explore effects of individual differences in Type A behavior pattern. Type A individuals are characterized by high levels of competitiveness, impatience, need for achievement, and a feeling of being constantly under pressure. They tend to experience higher levels of stress than do Type B individuals (Caplan & Jones, 1975; Ivancevich & Matteson, 1984; Motowidlo, Packard, & Manning, 1986), in part, apparently, because they react more strongly to stressful situations. Therefore, we expect to find that Type A score is nega-

This research was supported by a grant from the Center for Research in the College of Business Administration at the Pennsylvania State University.

Thanks are due to Bob Griffin and his staff at the Behavioral Learning Center for their help in producing the videotape that accompanied the inbasket. We also thank Hank Sims and Bob Sinclair for helpful comments on an earlier draft.

Correspondence concerning this article should be addressed to Stephan J. Motowidlo, Personnel Decisions Research Institute, 43 Main Street Southeast, Suite 405, Minneapolis, Minnesota 55414.

tively related to the favorability and dispersion of performance ratings and also that it interacts with stressor effects such that this negative relation is stronger in the more stressful experimental condition.

Method

Sample

Participants consisted of 120 undergraduate students, juniors and seniors in business administration, who were randomly assigned to one of four experimental conditions in a 2×2 crossed design. They were paid \$6 for taking part in the study. There were equal numbers of men and women in the four groups.

Design

Participants were asked to assume the role of a sales manager and complete an inbasket lasting for 45 min. There were two versions of the inbasket exercise, one much more difficult with greater information processing demands (high work load) than the other (low work load). In this way, we manipulated two levels of experimentally induced stress.

Participants in the high-work-load group were asked to go through a complicated inbasket exercise while being interrupted frequently with additional information presented on a videotape. The inbasket material for this group consisted of problems involving interdepartmental conflicts, production delays, supervisor-subordinate problems, and general information concerning the routine operations of the company. The videotape interruptions depicted visits to the sales manager's office by his or her superior, subordinates, and friends who provided additional information that usually involved material in the inbasket. There were 11 interruptions of about 45 s each.

Participants in the low-work-load condition completed a less complicated version of the inbasket with problem situations that were more routine. They did not see the interruption videotape while working on the inbasket.

The second experimentally manipulated independent variable was order of information presentation. One half of the participants in the high- and low-work-load conditions saw a performance videotape before they worked on the inbasket, and the other half saw it afterward. Accordingly, two levels of work load were crossed with two levels of order of information presentation in this experimental design.

Procedures

All participants were given a brief introduction to the study and were asked to sign informed consent forms. Next, the experimenter took each person's pulse at the wrist by counting the number of beats per 15 s. They then filled out the student version (Form T; Matthews, 1982) of the Jenkins Activity Survey (Jenkins, Rosenman, & Zyzanski, 1974), which measures Type A behavior pattern. Its internal consistency reliability in this sample was .70.

One half of the participants in the high-work-load group and one half in the low-work-load group observed the performance videotape *before* working on the inbasket exercises. The performance videotape lasted for 8 min and showed a manager dealing with his subordinate in an appraisal interview. The videotape and the accompanying rating scales used here were adapted from those developed by Borman, Hough, and Dunnette (1976). We used only one of the performance episodes developed by Borman et al., and we abbreviated the directions and some of the anchors on the rating scales so that ratees would not have to read so much material to use them. Next, the participants worked on the inbasket exercise for 45 min. Another measure of their pulse was obtained to provide a physiological measure of stress reaction (Fried, Rowland,

& Ferris, 1984). Then they filled out the short form of the Multiple Affect Adjective Checklist (MAACL; Zuckerman & Lubin, 1965), which served as a measure of their mood state. Participants were asked to indicate how they "feel at present," as a measure of temporary mood states. Checklist items include adjectives such as *calm*, *afraid*, *understanding*, and *nervous*. The participants were asked to place a check mark on those items that described how they were feeling. The MAACL was scored for anxiety, depression, and hostility. The internal consistencies of the three scales in this sample are .68, .63, and .69, respectively.

Participants also filled out a 15-item questionnaire designed to measure subjective stress. Three examples of items follow: (a) "I was overwhelmed by all the information that was presented in the inbasket," (b) "It was difficult to concentrate while going through the inbasket," and (c) "I felt very tense while going through the inbasket." The internal consistency reliability of the measure of subjective stress in this sample was .83. Finally, participants rated the performance of the manager in the performance videotape on seven scales adapted from Borman et al. (1976) and on an eighth scale of overall effectiveness.

The other half of the high- and low-work-load groups saw the performance videotape *after* they worked on the inbasket exercises. For them, the sequence of procedures after completing the Type A scale was as follows: inbasket exercise, pulse measurement, MAACL, stress questionnaire, performance videotape, and rating scales related to the performance videotape.

Dependent Variables

Ratings obtained from the participants yielded scores for favorability and dispersion. Favorability was computed simply as the sum of the eight rating scales. A dispersion score for each participant was computed as the squared difference between each of the seven behaviorally anchored scales adapted from Borman et al. (1976), and their mean, summed over the seven scales. This represents the variability or dispersion of a rater's ratings on the seven performance dimensions around their mean.

Results

A multivariate analysis of variance (MANOVA) with hostility, depression, anxiety, subjective stress, and pulse rate as dependent variables was conducted as a manipulation check. Sex, order of information presentation, and work load were the independent variables, and pulse rate measured before the experimental manipulation was a covariate. There was a significant multivariate effect of work load, $F(5, 107) = 13.35, p < .01$. Multivariate effects of sex, $F < 1.0$; order of information presentation, $F(5, 107) = 1.62, ns$; and the interaction terms, all $Fs < 1.28, ns$, were not significant. Univariate tests showed significant effects of work load on pulse, $F(1, 111) = 8.00, p < .01, \omega^2 = .055$; anxiety, $F(1, 111) = 27.32, p < .01, \omega^2 = .185$; hostility, $F(1, 111) = 10.19, p < .01, \omega^2 = .072$; depression, $F(1, 111) = 4.82, p < .05, \omega^2 = .032$; and subjective stress, $F(1, 111) = 61.15, p < .01, \omega^2 = .336$. The pattern of means indicates that participants in the high-work-load condition had more rapid pulse rates and reported stronger feelings of anxiety, hostility, depression, and subjective stress. These results suggest that the manipulation successfully aroused more stress among persons in the high-work-load condition than among persons in the low-work-load condition.

Another three-way MANOVA was conducted with favorability and dispersion as dependent variables. There was a significant multivariate effect for work load, $F(2, 111) = 3.84, p < .05$, but

Table 1
Means and Standard Deviations According to Experimental Condition

	Performance information		
Dependent variable	Before inbasket	After inbasket	Overall <i>M</i>
High-work-load condition			
Favorability			
<i>M</i>	38.97	38.80	38.88
<i>SD</i>	6.13	7.57	6.83
Dispersion			
<i>M</i>	8.59	9.04	8.81
<i>SD</i>	4.13	5.31	4.72
Low-work-load condition			
Favorability			
<i>M</i>	40.47	38.17	39.32
<i>SD</i>	6.13	5.47	5.88
Dispersion			
<i>M</i>	11.62	11.54	11.58
<i>SD</i>	7.10	7.08	7.03

not for order of information presentation, $F < 1.0$; sex, $F < 1.0$; or any of the interactions, all F s < 1.52 , *ns*. Univariate tests indicated a significant effect of the work load manipulation on dispersion, $F(1, 112) = 6.41$, $p < .05$, $\omega^2 = .043$, but not on favorability, $F < 1.0$. As shown in Table 1, ratings made by people who completed the more stressful inbasket showed less dispersion across performance dimensions than did ratings made by people who completed the less stressful inbasket.

Correlations between the eight rating scales were computed separately for persons in the high- and low-work-load conditions to explore further the effects of stress on dispersion. As shown in Table 2, 24 of the 28 correlations are stronger in the high-work-load condition. According to the nonparametric sign test (Siegel, 1956), this result is significant at $p < .01$, and lends additional support to the conclusion that people who have recently been stressed discriminate less between different dimensions when rating job performance.

Hierarchical regression analyses tested main and interaction

effects of Type A pattern on favorability and dispersion. Sex, order of information presentation, and their interaction terms were dropped from these analyses because they had no significant effects in earlier analyses. Work load was dummy coded as 0 (low) or 1 (high). Favorability and dispersion were regressed on independent variables and interaction terms in this order: (a) work load, (b) Type A score, and (c) work load by Type A score. At each step, the increase in R^2 and total R^2 were calculated and tested for significance.

Table 3 displays the results. As shown, none of the main or interaction effects contributes significantly to explained variation in favorability, and its final total R^2 is only .015, *ns*. Dispersion, however, is another matter. Work load alone explains 5.1% ($p < .05$) of the variation in dispersion scores. The main effect for Type A scores does not add significantly to this explained variation, but the interaction of work load and Type A score explains 5.5% ($p < .01$) of the variance beyond that explained by the two main effects. All together, the main effects and interaction explain a total of 12.3% ($p < .01$) of the variance in dispersion.

The final regression equation for dispersion is $Y = 6.20 + 4.34 A + 0.61 B - 0.83 AB$, where A stands for work load and B stands for Type A score. To examine the nature of the significant interaction of work load and Type A score, we computed regression equations for effects of Type A on dispersion separately for the low- and high-work-load conditions. This was accomplished by substituting the dummy codes for work load in the full regression equation just stated. The two regression equations that resulted follow: (a) for the low-work-load condition, $Y = 6.20 + 0.61B$, and (b) for the high-work-load condition, $Y = 10.54 - 0.22B$. Then, for B in each equation, we substituted the value of the Type A score one standard deviation above the mean (11.77) and one standard deviation below the mean (4.81) and plotted the two regression lines in Figure 1.

According to our hypotheses: (a) The regression line in the high-work-load condition should be below the one in the low-work-load condition (the predicted main effect of work load); (b) the difference between high- and low-work-load conditions should increase with increasing Type A score (the predicted interaction of Type A score and work-load condition); and (c) both slopes should be negative (the predicted main effect of Type A pattern). Two of these predictions are supported by the

Table 2
Correlations Between Rating Scales Separately Computed for High- and Low-Work-load Conditions

Scale	1	2	3	4	5	6	7	8
1	—	.23	.24	.06	.33*	.27*	.16	.43*
2	.40**	—	.20	.22	.03	.27*	.05	.41**
3	.28*	.36**	—	.26*	.18	.27*	.21	.40**
4	.43**	.47**	.56**	—	.29*	.22	.23	.43**
5	.37**	.46**	.30*	.38**	—	.36**	.19	.37**
6	.13	.37**	.32*	.34**	.25	—	.32*	.47**
7	.29*	.42**	.22	.42**	.36**	.27*	—	.26*
8	.51**	.58**	.43**	.54**	.50**	.43**	.46**	—

Note. Correlations above the diagonal are from the low-work-load condition, and correlations below the diagonal are from the high-work-load condition.

* $p < .05$ (two-tailed). ** $p < .01$ (two-tailed).

Table 3
Hierarchical Regression Analysis of Effects on Favorability and Dispersion

Variables added to the regression equation	Effects on favorability		Effects on dispersion	
	Increase in R^2	Total R^2	Increase in R^2	Total R^2
Work load (A)	.001	.001	.051*	.051*
Type A Score (B)	.000	.001	.017	.069*
A \times B	.014	.015	.055**	.123**

* $p < .05$. ** $p < .01$.

data, but the third is not. According to results presented in Figure 1: (a) For most of its range in Type A scores, the regression line in the high-work-load condition is below the one in the low-workload condition; (b) the difference between high- and low-workload conditions increases with increasing Type A scores; and (c) the slope in the high-work-load condition is slightly negative, but the slope in the low-work-load condition is markedly positive. This third point disconfirms our prediction of a negative relation between Type A score and dispersion. There is no significant main effect of Type A score, and the interaction is such that in the low-work-load condition, Type A score is positively instead of negatively related to dispersion. The correlation between Type A score and dispersion in the low-work-load condition is .32 ($p < .05$), and in the high-work-load conditions it is $-.15$, *ns*. These results indicate that the Type A pattern interacts with work load in affecting the dispersion of performance ratings, but not in the manner we expected.

Discussion

Results of this study show that stressful experiences affect the way people process information to make evaluative judgments. Participants who completed a stressful inbasket simulation subsequently produced performance ratings that were less differentiated across dimensions. Their ratings were also more highly intercorrelated. These results support the hypothesis derived from S. Cohen's (1980) psychic cost explanation that stressful experiences consume some portion of attentional capacity that could otherwise be devoted to the task of forming performance judgments. As a result, raters depend more heavily on their first impressions and differentiate less between performance dimensions.

Whether the performance information is presented before or after the stressful experience does not make a great deal of difference. The univariate interaction effect of work load and order of information presentation on dispersion was not significant at $p < .05$. (According to J. Cohen, 1977, the statistical power available in this sample for detecting a medium-size interaction effect, with alpha set at .05, is approximately 78%.) Thus, the main effect of work load on dispersion does not differ significantly in the two conditions of information order. Because there is no evidence that the main effect of work load on dispersion varies according to order of information presentation, we conclude that a stressful experience causes a decrease

in dispersion both when it precedes and when it follows the presentation of performance information.

When the stressor follows the presentation of performance information, its effects on dispersion cannot be attributed to the input stage of information processing because by the time the stressful experience occurs, information has already been attended to and encoded. In that condition, therefore, effects of stress must be attributed to the retrieval stage. When the stressor precedes the presentation of performance information, however, its effects on dispersion can be attributed *either* to the input stage, the retrieval stage, or *both* input and retrieval stages. Consequently, of the three possibilities—that stress affects (a) input but not retrieval, (b) retrieval but not input, or (c) both input and retrieval—our results rule out the first possibility but not the other two. We conclude, therefore, that stressful experiences affect *at least* the retrieval stage of information processing and might affect the input stage as well.

We began this study expecting that Type A scores would be negatively related to dispersion and that the negative relation would be stronger in the high-work-load condition. This expectation was not confirmed. There was an interaction, but it took the form of a positive relation in the low-work-load condition and a weak (if any) negative relation in the high-work-load condition.

We did not find the predicted effects of stress on favorability. We expected that stress would produce a negative mood state that would bias attention and retrieval processes toward negative performance information. This prediction was based upon the assumption that raters "care" whether the ratee's performance is good or poor, so that positive rater moods would be associated with positive performance information and negative moods would be associated with negative information. In real work organizations, this assumption is quite reasonable. Super-

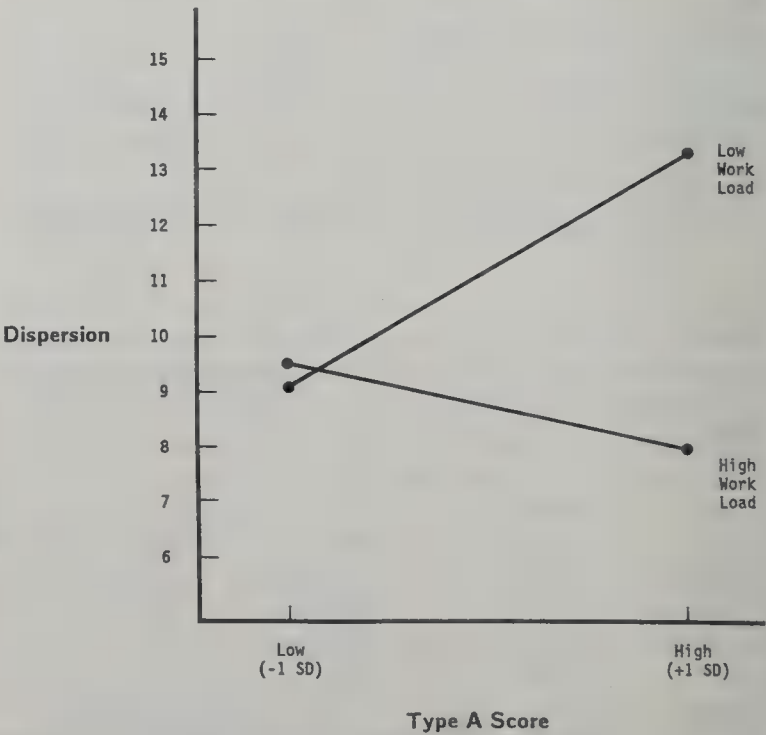


Figure 1. Regression of dispersion on Type A score according to work-load condition.

visors and managers are generally likely to prefer that their subordinates perform well. In this experiment, however, there was no reason for research participants to want the manager to do well. Because of their probable indifference toward the manager's performance, this study might not afford a good test of the hypothesis that raters evaluate performance less generously after they have been stressed.

We should note that order of information in this experiment was confounded with time delay between performance observation and evaluation. In the condition in which performance information was presented before the inbasket exercise, there was a delay of 45 min (while raters worked on the inbasket) before raters recorded performance evaluations. In the other condition, they worked on the inbasket first, then saw the performance videotape, and then rated the performance with no delay. However, because there were no main or interaction effects of order of information presentation, this confound does not seem to present a serious problem here.

Several steps were taken to enhance the likelihood that results of this study would generalize to management populations. The research participants were undergraduate majors in business administration, most of whom aspire to careers in business and management. The independent variable, stress, was manipulated in the form of a carefully developed inbasket exercise designed to simulate managerial working conditions as realistically as the constraints of a laboratory experiment would permit. And the dependent variable, performance evaluation, was measured with another management simulation—a videotaped portrayal of a manager dealing with a subordinate. However, because raters were exposed only to a brief encounter with the ratee, this study may be limited to first-impression effects that might not generalize to situations where raters have long-term exposure to the ratee.

If managers are stressed by their jobs in ways similar to the ways in which our research participants were stressed by the complex inbasket simulation, they are also likely to be vulnerable to similar effects on information processing. When required to make judgments of performance effectiveness, or any other multidimensional judgments related to personnel decisions, they too might not discriminate as much as they probably should between different dimensions of behavior. As a result,

they too might make judgments that are probably too heavily colored by global evaluative impressions.

References

- Borman, W. C., Hough, L. M., & Dunnette, M. D. (1976). *Performance ratings: An investigation of reliability, accuracy and relationships between individual differences and rater error*. Minneapolis, MN: Personnel Decisions, Inc.
- Bower, G. H. (1981). Mood and memory. *American Psychologist*, 36, 129-148.
- Caplan, R. D., & Jones, K. W. (1975). Effects of work load, role ambiguity, and Type A personality on anxiety, depression, and heart rate. *Journal of Applied Psychology*, 60, 713-719.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.
- Cohen, S. (1980). Aftereffects of stress on human performance and social behavior: A review of research and theory. *Psychological Bulletin*, 88, 82-108.
- Fried, Y., Rowland, K. M., & Ferris, G. R. (1984). The physiological measurement of work stress: A critique. *Personnel Psychology*, 37, 583-615.
- Isen, A. M., Shaker, T. E., Clark, M., & Karp, L. (1978). Affect, accessibility of material in memory, and behavior: A cognitive loop? *Journal of Personality and Social Psychology*, 36, 1-12.
- Ivancevich, J. M., & Matteson, M. T. (1984). A Type A-B person-work environment interaction model for examining occupational stress and consequences. *Human Relations*, 37, 491-513.
- Jenkins, C. D., Rosenman, R. H., & Zyzanski, S. J. (1974). Prediction of clinical coronary heart disease by a test for the coronary-prone behavior pattern. *New England Journal of Medicine*, 23, 1271-1275.
- Matthews, K. A. (1982). Psychological perspectives on the Type A behavior pattern. *Psychological Bulletin*, 91, 293-323.
- Motowidlo, S. J., Packard, J. S., & Manning, M. R. (1986). Occupational stress: Its causes and consequences for job performance. *Journal of Applied Psychology*, 71, 618-629.
- Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.
- Zuckerman, M., & Lubin, B. (1965). *Manual for the Multiple Affect Adjective Checklist*. San Diego, CA: Educational and Industrial Testing Service.

Received March 25, 1986

Revision received September 9, 1986

Accepted December 1, 1986 ■

The Systematic Distortion Hypothesis, Halo, and Accuracy: An Individual-Level Analysis

Steve W. J. Kozlowski and Michael P. Kirsch
Michigan State University

In several social perception studies investigators have concluded that raters' semantic conceptual similarity schemata serve to guide and constrain dimensional covariance in the rating judgment process. This effect has been hypothesized to be most likely when ratings are memory based and raters lack relevant job or ratee information. Recent research that has explored the effects of conceptual similarity schemata on performance ratings and halo error has provided some limited support for this systematic distortion hypothesis (SDH). However, these studies are limited because researchers have examined this phenomena using group-level analyses, whereas the theory references individual-level judgment processes. The present study investigated the phenomena at the individual level. The effects of varying levels of rater job knowledge (high, medium, and low) and familiarity with ratees (high and low) were examined for conceptual similarity-rating and rating-true-score covariation relations, for measures of halo, and for rating accuracy components. Results provided support for the SDH, but indicated a boundary condition for its operation and revealed some surprising findings for individual-level rater halo.

In a series of studies, Shweder and D'Andrade (D'Andrade, 1974; Shweder, 1975, 1980, 1982; Shweder & D'Andrade, 1980) examined whether raters' conceptual similarities among rating categories provide schemata for dimensional covariation. They concluded that ratings were more sensitive to raters' implicit conceptual similarities than to actual behavioral covariation. Newcomb (1931) also noted higher intercorrelations among ratings than among behavioral frequency measures for the same set of ratees. He suggested that halo error may result from raters' implicit covariance assumptions that overestimate true behavioral covariation.

Drawing on this work examining personality traits, Cooper (1981a; 1981b) posited that raters' illusory covariance assumptions may cause halo error in performance ratings. That is, although there is some level of veridical performance covariation (true halo), halo error results when rating covariation exceeds true halo (Bingham, 1939). Cooper (1981a) suggested that raters mistake the conceptual similarity (CS; Shweder, 1982) between performance dimension labels for true behavioral covariance. To the extent that raters' CS schemata overstate actual performance covariation and CS schemata influence the pattern of rating covariation, CS is implicated in halo error.

It has been hypothesized that CS schemata influence rating covariation when ratings are obtained under *difficult memory*

conditions and contain 20%-30% error variance (Shweder, 1980, 1982), that is, when decayed observations are recalled from memory and raters lack knowledge of the performance domain or familiarity with the ratee, or both (Borman, 1983). The systematic distortion hypothesis (SDH) asserts that raters' recall processes are systematically biased in the direction of their implicit covariance schemata when these conditions are present. Thus, as job and ratee knowledge for memory-based ratings decrease, there are expected increases in the CS-rating covariance association, with corresponding increases in halo error and inaccuracy.

Cooper (1981a) tested the SDH in the performance-appraisal domain. A sample of 10 persons provided CS judgments, and data obtained from Borman (1977, 1979) provided performance ratings based on observations of videotaped behaviors and performance true scores. Cooper correlated parallel rating intercorrelations, true-score intercorrelations, and CS judgments. He found that raters were more sensitive to actual behavior than to conceptual similarity, in contradiction to the SDH. As Cooper noted, however, his study was not an adequate test of the SDH because respondents assigned ratings immediately after viewing the tapes and did not provide sufficient memory decay to facilitate the use of CS schemata (cf. Murphy & Balzer, 1986; Shweder, 1980, 1982).

Kozlowski, Kirsch, and Chao (1986) identified other difficulties with this study. First, the CS judgments were derived from a small, independent sample of raters. To adequately test the SDH, the same raters should have provided both the CS judgments and the ratings. Second, and more important, the videotape methodology that was used derived true-score estimates from a set of pooled rater judgments. Thus, any pervasive cognitive distortion process affecting ratings would also be expressed in the true scores.

Kozlowski et al. (1986) rectified these limitations by using

Portions of this research were presented at the First Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, 1986.

We gratefully acknowledge two anonymous reviewers whose thoughtful comments improved this manuscript, and Scott Cohen for his assistance on parts of the data analyses.

Correspondence concerning this article should be addressed to Steve W. J. Kozlowski, Department of Psychology, Michigan State University, East Lansing, Michigan 48824-1117.

baseball players as target ratees. These ratees provided objective behavioral frequency true scores uncontaminated by rater judgment processes. The study examined the effects of job (baseball) knowledge and ratee familiarity on the relation between CS and rating covariance. A low CS–rating correlation was found for high-job-knowledge raters rating high-familiarity players, compared to low-job-knowledge raters rating high-familiarity players. In addition, the CS–rating relation dramatically increased when the high-knowledge raters rated low-familiarity players. Finally, halo systematically increased as job and ratee knowledge decreased.

The results were interpreted cautiously. It was noted that the hypothetical cognitive process underlying the SDH and its effect on halo error references the individual level, whereas the research has used group-level data to generate the covariance matrices. Aggregating data over individuals reduces the effects of individual variation and may increase the apparent magnitude of the CS effect (cf. L. James, 1982). Thus, the impact of individual-level CS on individual-level rating covariation, halo, and accuracy remains to be addressed.

The present study was designed to address these issues. Consistent with the SDH, we hypothesized that the relation between CS covariance and rating covariance would be higher, whereas the relation between rating covariance and true-score covariance would be lower, when raters lacked relevant job knowledge. That is, when raters have little concrete experience in the relevant job domain they are more likely to rely on their implicit similarities among rating dimension labels to guide the pattern of ratings. As this process occurs, there should be corresponding increases in halo error and rating inaccuracy (Borman, 1983).

The SDH predicts that the same pattern would hold and increase when raters lack familiarity with ratees. These relations may hold, however, only for raters with sufficient job knowledge. Kozlowski et al. (1986) reported a decline in the CS–rating-covariance relation for low-job-knowledge raters evaluating unfamiliar versus familiar ratees. Although the difference was not significant, it implied a possible boundary condition for the SDH.

Romer and Revelle (1984) indicated that semantic labels for aspects of performance are implicated during the encoding of observations and thereby also serve as schemata for recall processes (cf. Nathan & Lord, 1983). Thus, the CS schemata of raters with more observational experience in the relevant performance domain (job knowledge) are likely to be more concrete, salient, and accessible. They are able to observe a given performance segment and reliably encode it according to the performance dimension labels it references. Such raters are likely to have access to specific performance information for well-known ratees. With specific information serving as reference, a CS schema would guide ratings for performance dimensions on which the rater lacked more detailed information. As the level of familiarity with the ratee decreased and corresponding performance information decreased, the CS schema would tend to serve a larger role in the pattern of rating responses. Low-job-knowledge raters' CS schemata are likely to be more abstract and ill defined. Their ability to reliably encode an observed performance segment with respect to its relevant performance dimensions would be lower. With less detailed and more tentative information about the performance of specific famil-

iar ratees available, CS would play a stronger role in the pattern of ratings. When rating unfamiliar ratees, however, there may be no specific information accessible and no guidance for determining the general level of the ratings. Without some reference for anchoring the level of the ratings, the CS schema alone may not be sufficient to affect the pattern of responses. If so, the result would be ratings that appear unstructured. Decrements in the CS–rating-covariation linkage, decreased rating intercorrelations, increased within-ratee variance, and decreased accuracy would be expected.

Method

Sample

The respondents were undergraduate students at a large midwestern university who participated in return for class credit ($N = 315$). Completed questionnaires were returned by 296 of the participants. Because of methodological difficulties associated with the use of individual-level correlation matrices, 69 cases were excluded from the analysis, making the sample size 227. This problem is discussed later.

Procedure

Systematic distortions biased in the direction of CS schemata are hypothesized to occur under difficult memory conditions (i.e., when ratings are memory based and sufficient time between observation and recall have occurred to allow memory decay; Shweder, 1980). To ensure that the ratings were memory based and decayed, they were collected 3 weeks after the end of the previous (1984) baseball season. Instructions to the raters indicated that they were to rate player performance for the entire season. Thus, rating judgments had to aggregate relevant memory-based information over a 6-month period. This ensured difficult memory conditions.

The hypotheses proposed that the effects of CS schemata on covariation relations, halo, and accuracy would vary systematically by the level of job knowledge and ratee familiarity. These two conditions were created as follows.

Job knowledge. A job (baseball) knowledge quiz, developed by several subject matter experts, was piloted on an independent sample ($n = 53$). Ten items that best discriminated between high and low self-rated baseball knowledge were retained for the quiz. Respondents in the main study took the quiz and also made a self-rating of baseball knowledge on a 7-point scale.¹ Responses to both measures were normally distributed and correlated .73. Coefficient alpha reliability for the knowledge quiz was .77. The main sample was split into approximate thirds based on the quiz scores (high = 6 to 10 correct, $n = 67$; medium = 3 to 5 correct, $n = 89$; low = 0 to 2 correct, $n = 58$). The sample was split into thirds to better establish the boundary conditions for the SDH under the low-ratee-familiarity condition.

Ratee familiarity. An independent sample ($n = 79$) rated their familiarity with 90 baseball players on a 7-point scale, where higher scale values indicated greater familiarity. From this group of players, 10 each were selected as representing the extreme of the high- and low-familiarity ratings. Interrater reliability for the selected ratees, estimated by an intraclass correlation coefficient, was .99, indicating good agreement on ratee familiarity. Participants in the present study also rated their

¹ Anchors for the job-knowledge rating were as follows: 1 = *not very knowledgeable*—I seldom or never read about or watch baseball; 4 = *somewhat knowledgeable*—I sometimes follow the sport in newspapers and watch it on TV; and, 7 = *extremely knowledgeable*—I follow the sport in the newspaper and watch it on TV regularly.

familiarity with each ratee as a validity check on the preselection assignments.

Performance dimensions. The dimensions of baseball player performance used in the study were selected to be observable, important to success, and variable in terms of their level of intercorrelation. The 10 dimensions measured included batting average, and number of home runs, extra-base hits, runs batted in, runs scored, strikeouts, stolen bases, walks, fielding errors, and runs created.² Objective measures of these performance dimensions for the ratees in the study were obtained directly from official major league sources.³

Conceptual similarity judgments. To assess CS schemata, raters were asked to indicate the degree of similarity (0%–100% similar) for all unique pairings of the 10 performance dimensions. This resulted in 45 CS judgments. Instructions were given that explained the construct of CS and provided some examples. Definitions of the performance dimensions were also given.

Ratings and true scores. Each participant rated 10 ratees predefined as high in familiarity and 10 ratees predefined as low in familiarity, with all 20 ratees randomly sequenced to preclude order effects. Each of the players was rated on the 10 dimensions using 7-point scales. Scale anchors were matched to the content of each dimension with 1 indicating low performance and 7 indicating high performance. True scores for each ratee corresponding to the performance dimensions were matched with each participant's set of ratings. Because no objective data for 1984 were available for 3 of the low-familiarity players, these ratees were excluded from the analysis. This means that indices of covariation, halo, and accuracy for low-familiarity ratees were based on a set of 7 ratees; high-familiarity-ratee indices were based on a set of 10 ratees.

Results

Verification of Design Parameters

The design for the study included three levels of job knowledge and two levels of ratee familiarity, with repeated measures on the ratee-familiarity factor. To ensure that the selection criterion for rater assignment to the job-knowledge conditions was veridical, an analysis of variance (ANOVA) on the knowledge quiz scores was conducted. The results were significant, $F(2, 226) = 660.45, p < .05$. A follow-up Scheffé contrast showed that the groups were all significantly different ($p < .05$).

Participants' rated familiarity with the ratees was also examined to ensure that familiarity was not confounded with the job-knowledge conditions. Familiarity ratings were averaged for the high- and low-familiarity conditions. A repeated measures ANOVA indicated a nonsignificant job-knowledge effect, $F(2, 221) = .415, p > .05$; a significant ratee-familiarity effect, $F(1, 221) = 472.395, p < .05$; and a significant interaction, $F(2, 221) = 8.094, p < .05$. Inspection of the means showed transposition of the order of the means for job-knowledge groups across ratee-familiarity conditions. The means were ordered from high to low job knowledge for the high-familiarity condition and from low to high, though lower overall, for the low-familiarity condition. Scheffé contrasts for the job-knowledge groups in both ratee-familiarity conditions were not significant ($p > .05$).

To ensure that high and low familiarity were not confounded with ratee performance, a multivariate ANOVA was conducted on the ratees' true scores. The overall multivariate test, $F(10, 6) = 1.455, p > .05$, and all univariate tests were not significant ($p > .05$). The results of these validity checks indicated that the design parameters were valid.

Covariation

An ANOVA was conducted on the average CS-judgment values to ensure that the average levels of implicit dimensional covariance were equivalent across the job-knowledge conditions. The ANOVA was not significant, $F(2, 225) = .242, p > .05$.

To address the SDH, individual-level intercorrelation matrices were computed for the ratings and true scores. One half of each respondent's rating and true-score matrices, excluding the diagonals, were arrayed as vectors corresponding to the 45 CS judgments. All correlation values were transformed to z scores before analysis throughout the study and will not hereafter be specifically noted. Individual-level correlations were obtained to assess the covariation relations for the CS-rating and rating-true score vectors. The correlations were computed separately for the high- and low-ratee-familiarity groups.⁴

The covariation relation values were submitted to a repeated measures multivariate ANOVA. The results were significant for job knowledge, $F(4, 420) = 22.368, p < .001$; ratee familiarity, $F(2, 210) = 14.889, p < .001$; and the interaction term, $F(4, 420) = 9.218, p < .001$. Follow-up univariate ANOVAs were performed. For the CS-rating covariation relation, job knowledge was marginal, $F(2, 211) = 2.597, p = .08$, but both ratee familiarity, $F(1, 211) = 27.037, p < .001$, and the interaction term, $F(2, 211) = 13.923, p < .001$, were significant. For the rating-true-score covariation relation, job knowledge had a significant effect, $F(2, 211) = 38.771, p < .001$, whereas ratee familiarity, $F(1, 211) = .317, p > .05$, and the interaction term, $F(2, 211) = .571, p > .05$, did not. The pattern of rating-true-score covariation was relatively constant across ratee familiarity conditions, but was stronger for the high-job-knowledge group. Table 1 displays mean individual-level intermatrix z values and 95% confidence intervals for the six cells in the design. Mean correlation values were virtually identical.

Results consistent with the SDH were obtained for the CS-rating and rating-true-score covariation relations when raters were rating high-familiarity ratees. The relation between CS and rating covariation increased as job knowledge decreased.

² Runs created (RC; B. James, 1982) is an index of the player's ability to increase his team's score through all of the elements of his batting record. It was computed using the following formula:

$$RC = \frac{(\text{Hits} + \text{Walks})(\text{Total Bases})}{(\text{At Bats} + \text{Walks})}$$

³ The official sources were Sports Information Center, Quincy, Massachusetts (American League), and Elias Sports Bureau, New York, New York (National League).

⁴ Dimensional intercorrelations could not be computed when one or more of the dimensions lacked variance over ratees. For raters with one such dimension, each rater's specific ratee means were substituted for ratings on the dimension in question. To ensure that this procedure did not distort the results, the covariation analysis of variance was conducted with all of these cases removed. There were no meaningful differences in the results when these cases were included and excluded. Raters who failed to vary responses on more than one dimension were deleted from the analyses. An examination of the number of deleted cases (inasmuch as the uncomputable correlations would be implicated in increased covariance and halo) by job-knowledge conditions was not significant, $F(2, 66) = .336, p > .05$.

Table 1

Means, Standard Deviations, and 95% Confidence Intervals of Relations for Conceptual Similarity (CS), Rating (R), and True-Score (TS) Covariance

Job knowledge	Ratee familiarity					
	High			Low		
	<i>M</i>	<i>SD</i>	95% confidence interval	<i>M</i>	<i>SD</i>	95% confidence interval
High						
CS-R	.233	.209	.182 to .284	.272	.247	.212 to .332
R-TS	.340	.166	.299 to .380	.308	.254	.246 to .370
Medium						
CS-R	.261	.221	.214 to .308	.139	.202	.096 to .181
R-TS	.137	.175	.100 to .174	.148	.205	.105 to .191
Low						
CS-R	.344	.223	.285 to .402	.181	.197	.129 to .233
R-TS	.129	.136	.094 to .165	.115	.191	.064 to .165

Note. Intermatrix relations expressed as Fisher *r* to *z* values.

Moreover, the rating-true-score covariation relation was higher than the CS-rating covariation relation for high-job-knowledge raters, whereas this pattern was reversed for medium- and low-job-knowledge raters. This pattern of results, however, was not maintained when raters were rating less familiar ratees. The CS-rating covariation relation increased slightly for high-job-knowledge raters, but dropped significantly for medium- and low-job-knowledge raters. Thus, the boundary condition noted in the previous study with group-level data (Kozlowski et al., 1986) was also manifest in these data at the individual level. This phenomenon will be considered in more detail later.

Halo

The effects on halo error were investigated next. The measure of halo error was obtained by subtracting each respondent's average rating intercorrelation from the average true-score intercorrelation (Pulakos, Schmitt, & Ostroff, 1986)—what Bingham (1939) has termed *illusory halo*. Positive values indicated overstatement of dimensional covariation, whereas negative values indicated an understatement. Means for the illusory halo measure for the high-familiarity condition were $-.29$, $-.25$, and $-.22$, whereas low-familiarity means were $-.17$, $-.38$, and $-.34$ for high to low job knowledge, respectively. It was apparent that there was no halo error occurring. That is, all cell means indicated underestimates of covariation in the ratings relative to the true scores. To explore this further, a repeated measures ANOVA using the average rating intercorrelation was performed. This alternative measure of halo was used because it is more conventional and correlates perfectly with the illusory halo measure. The results indicated significant job knowledge, $F(2, 211) = 4.308$, $p < .05$; nonsignificant ratee familiarity, $F(1, 211) = .571$, $p > .05$; and significant interaction, $F(2, 211) = 16.525$, $p < .001$, effects. Table 2 shows the means, standard deviations, and 95% confidence intervals for the individual-level halo measure.

These results are most easily interpreted with reference to the average level of true-score intercorrelation, which was $.40$

($z = .425$) for high-familiarity ratees and $.45$ ($z = .489$) for low-familiarity ratees. The pattern of means for the high-ratee-familiarity condition is consistent with most previous research showing increasing halo with decreasing job knowledge. It indicates greater dimensional differentiation by higher job-knowledge raters, although all of the job-knowledge groups underestimated the true level of dimensional covariation. When rating low-familiarity ratees, high-job-knowledge raters exhibited significant increases in their average rating intercorrelation, although they still underestimated the true level of covariation. Medium- and low-job-knowledge raters, however, showed small nonsignificant decreases in their rating intercorrelations.

This result implied greater within-ratee variance across dimensions for lower job-knowledge raters evaluating unfamiliar ratees. An examination of raters' standardized within-ratee standard deviations (SSD; Pulakos et al., 1986) confirmed this. Familiar ratee mean SSDs were $.864$, $.847$, and $.838$, whereas unfamiliar mean SSDs were $.797$, $.883$, and $.877$, for high- to low-job-knowledge groups, respectively.

Accuracy

The effects of job knowledge and ratee familiarity on four measures of rating accuracy—differential, stereotypic, overall, and correlational—were addressed (cf. Cronbach, 1955; Wiggins, 1973). Differential accuracy reflects the degree to which a rater is sensitive to differences among ratees on each of the performance dimensions. Stereotypic accuracy reflects a rater's sensitivity to the normative level of ratee performance. Overall accuracy reflects the degree to which ratings are similar in magnitude to the true scores. And correlational accuracy reflects the degree to which the rating profiles are sensitive to the true-score profiles of the ratees. These indices are not independent, as both differential and stereotypic accuracy are components of correlational and overall accuracy (Cronbach, 1955).

To operationally define the accuracy measures, the rating and true-score metrics were standardized within raters for the high- and low-familiarity groups, separately. Computational formu-

Table 2
Means, Standard Deviations, and 95% Confidence Intervals for the Average Rating Intercorrelation Halo Measure

Job knowledge	Ratee familiarity					
	High			Low		
	M	SD	95% confidence interval	M	SD	95% confidence interval
High	.134	.127	.103 to .165	.321	.422	.218 to .424
Medium	.176	.174	.140 to .213	.108	.167	.073 to .143
Low	.203	.214	.146 to .259	.147	.200	.095 to .200

Note. Average rating intercorrelation expressed as Fisher *r* to *z* values.

lae for differential, stereotypic, and overall accuracy were taken from Wiggins (1973). Correlational accuracy was obtained by correlating each rater’s ratings for a ratee with the true scores and then averaging over ratees. Smaller values for differential, stereotypic, and overall accuracy indicate greater accuracy, whereas larger values indicate greater correlational accuracy.

We hypothesized that more information would lead to greater accuracy. Thus, the highest accuracy was expected to occur for high-familiarity ratees, with accuracy decreasing across job-knowledge conditions. We expected that there would not be substantial differences across the job-knowledge conditions for low-familiarity ratees, although the accuracy of high-job-knowledge raters was generally expected to be higher.

A repeated measures multivariate ANOVA was conducted on the accuracy measures. There were significant interaction, $F(8, 364) = 6.990, p < .001$; job-knowledge, $F(8, 364) = 5.949, p < .001$; and ratee-familiarity effects, $F(4, 182) = 40.991, p < .001$. Univariate ANOVAs were conducted. All of the factors were significant ($p < .001$), with the exception of job knowledge for stereotypic accuracy. Results of these analyses are shown in Table 3. Cell means, standard deviations, and 95% confidence intervals for the accuracy measures are displayed in Table 4.

In sum, high-job-knowledge raters exhibited consistently superior accuracy for high-familiarity ratees. Indeed, the high-, medium-, and low-job-knowledge means were perfectly ordered, with high-job-knowledge raters significantly more accurate than medium- and low-job-knowledge raters across the four accuracy components. Medium- and low-job-knowledge raters’ accuracy was not significantly different for any of the components.

Table 3
Analysis of Variance Results for Accuracy Components

Design parameters	Accuracy component <i>F</i> values			
	Differential ^a	Stereotypic ^a	Correlational ^b	Overall ^a
Job knowledge	33.433*	1.707	12.299*	18.759*
Ratee familiarity	56.401*	45.517*	123.197*	102.992*
Interaction	9.313*	18.260*	22.581*	21.567*

^a *Df*s = 2, 211; 1, 211; and 2, 211, respectively.

^b *Df*s = 2, 198; 1, 198; and 2, 198, respectively.

* $p < .001$.

The results were less clear-cut for low-familiarity ratees. It was generally impossible to distinguish the job-knowledge groups, although high-job-knowledge raters had the best differential and the worst stereotypic accuracy. High-job-knowledge raters also exhibited consistently significant declines in accuracy across ratee-familiarity conditions. Medium-job-knowledge raters showed significant declines for the components, with the exception of stereotypic accuracy. Low-job-knowledge raters showed nonsignificant declines across ratee-familiarity conditions.

Correlational Analysis

To integrate the results for the covariation, halo, and accuracy analyses, measures of these outcomes were correlated. Intercorrelations were computed for high- and low-familiarity-ratee sets, separately, as shown in Table 5. Variables included job knowledge, treated as a continuous variable (baseball-knowledge quiz score); CS-rating and rating-true-score covariation; average rating intercorrelation and SSD (see Pulakos et al., 1986) measures of halo; and differential, stereotypic, correlational, and overall measures of accuracy.

For the high-ratee-familiarity condition, job knowledge had a significant negative correlation with CS-rating covariation. The greater the degree of job knowledge, the less CS and rating covariation were related. Job knowledge was also significantly related to rating-true-score covariation, indicating that greater job knowledge resulted in a better match between rating and true-score covariation profiles. This is consistent with the SDH. For the low-ratee-familiarity condition, job knowledge exhibited a significant positive correlation with CS-rating covariation. Job-knowledgeable raters exhibited a tendency to more closely match their pattern of rating covariation with their CS schemata. The association between job knowledge and rating-true-score covariation remained positive and significant, but dropped considerably compared with the high-ratee-familiarity condition. These outcomes for the low-familiarity condition are at variance with the SDH (cf. Shweder, 1980), but consistent with the interpretation that job knowledge may set boundary conditions for the operation of the SDH. Raters with lower job knowledge exhibited weak relations between their CS schemata and pattern of rating covariation when rating unfamiliar ratees.

Job knowledge was negatively correlated with both halo measures under the high-ratee-familiarity condition, but positively

Table 4
Means, Standard Deviations, and 95% Confidence Intervals for Differential, Stereotypic, Correlational, and Overall Accuracy

Job knowledge	Ratee familiarity					
	High			Low		
	<i>M</i>	<i>SD</i>	95% confidence interval	<i>M</i>	<i>SD</i>	95% confidence interval
Differential						
High	0.954	0.186	0.908 to 0.999	1.214	0.288	1.143 to 1.284
Medium	1.213	0.209	1.169 to 1.257	1.348	0.214	1.303 to 1.394
Low	1.258	0.193	1.208 to 1.309	1.297	0.195	1.246 to 1.348
Stereotypic						
High	0.326	0.163	0.286 to 0.366	0.628	0.268	0.563 to 0.693
Medium	0.419	0.170	0.383 to 0.455	0.457	0.198	0.415 to 0.499
Low	0.436	0.195	0.385 to 0.487	0.510	0.221	0.452 to 0.568
Correlations ^a						
High	0.246	0.133	0.211 to 0.281	−0.001	0.148	−0.040 to 0.038
Medium	0.078	0.100	0.055 to 0.098	−0.019	0.156	−0.053 to 0.015
Low	0.061	0.104	0.034 to 0.087	0.014	0.122	−0.017 to 0.046
Overall						
High	1.274	0.329	1.193 to 1.354	1.842	0.294	1.770 to 1.914
Medium	1.609	0.287	1.548 to 1.670	1.805	0.287	1.745 to 1.866
Low	1.656	0.266	1.586 to 1.726	1.807	0.308	1.726 to 1.888

Note. Smaller values indicate greater differential, stereotypic, and overall accuracy. Correlational accuracy expressed as Fisher *r* to *z* values.
^a Due to some uncomputable values, correlational accuracy samples were 60, 83, and 58 for high-, medium-, and low-job-knowledge groups, respectively.

correlated with them in the low-familiarity condition. When considered with the covariation relations noted previously, these results suggest support for the hypothesized link between CS–rating covariation and halo.
Job knowledge was significantly and positively related to the four measures of accuracy for the high-ratee-familiarity group, but exhibited little relation to accuracy for the low-familiarity

group. This is consistent with the significant ANOVA interaction observed previously.
Conceptual similarity–rating covariation was positively associated with halo, and rating–true-score covariation was not significantly related to halo, for the high-ratee-familiarity condition. For the low-familiarity condition, CS–rating covariation and halo correlations increased substantially. This result further

Table 5
Correlations Among Job Knowledge, Covariation, Halo, and Accuracy for High and Low Ratee Familiarity

Variable	Knowledge	Covariation		Halo		Accuracy			
	1	2	3	4	5	6 ^a	7 ^a	8 ^a	9
1. Job knowledge	1.00	−.17	.52	−.15	−.15	−.58	−.30	−.53	.68
2. Conceptual similarity–rating	.14	.38	.17	.22	−.21	−.01	−.13	−.09	−.06
3. Rating–true score	.29	.61	.31	.07	−.11	−.63	−.17	−.51	.63
4. Rating intercorrelation	.16	.38	.44	.16	−.89	−.30	.27	−.06	−.11
5. Standardized <i>SD</i>	−.19	−.41	−.44	−.90	.09	.34	−.21	.13	.05
6. Differential accuracy ^a	−.08	−.22	−.22	−.50	.48	.21	.29	.81	−.83
7. Stereotypic accuracy ^a	.15	.17	.20	.51	−.53	−.13	−.04	.74	−.36
8. Overall accuracy ^a	.05	−.04	−.01	.01	−.04	.66	.66	.06	−.73
9. Correlational accuracy	−.01	.06	.02	.06	−.07	−.79	−.10	−.68	.08

Note. Correlations for high ratee familiarity are above the diagonal (*n* = 206); low ratee familiarity correlations are below the diagonal (*n* = 202); diagonal values are correlations between familiarity conditions.
^a Metrics are reversed for these variables; low values indicate greater amounts of the variable.
p ≤ .05 for values ≥ .14; *p* ≤ .001 for values ≥ .22; two-tailed.

substantiates the relation between individual-level CS schemata and halo in ratings. The relation between rating–true-score covariation and halo also increased for the low-familiarity condition. This result probably reflects the increase in the average true-score intercorrelation for the ratees in the low-familiarity condition. Otherwise, these correlational results are consistent with the hypotheses.

Conceptual similarity–rating covariation was not significantly correlated with accuracy, whereas rating–true-score covariation was, for the high-familiarity condition. For the low-familiarity condition, CS–rating covariation and rating–true-score covariation were significantly and positively related to differential accuracy, but negatively related to stereotypic accuracy and not significantly related to overall and correlational accuracy.

Finally, both halo measures exhibited the same pattern of relations with accuracy across the ratee-familiarity conditions. Halo was positively and significantly correlated with differential accuracy, negatively correlated with stereotypic accuracy, and not significantly correlated with correlational and overall accuracy. The positive relation between halo and differential accuracy is consistent with previous research (cf. Cooper, 1981b; Murphy & Balzer, 1986).

Discussion

Covariation, Halo, Accuracy, and the Systematic Distortion Hypothesis Boundary

The results of this study provided support for the SDH at the individual level and further substantiated an apparent boundary condition for its operation. When raters evaluated familiar ratees—presumably possessing some memory-based information about the targets—those who were more knowledgeable about the performance domain exhibited significantly smaller CS–rating covariation and significantly larger rating–true-score covariation linkages than did less knowledgeable raters. Indeed, the covariation means were perfectly ordered by job knowledge in the hypothesized directions.

The results for halo and accuracy paralleled the covariation findings for the high-ratee-familiarity condition. More job-knowledgeable raters exhibited less halo and greater differential, stereotypic, overall, and correlational accuracy than did raters less knowledgeable about baseball. Again, the means were ordered by job knowledge in the expected directions. This pattern of results is consistent with the SDH (cf. Shweder, 1982) and conventional views regarding the effects of knowledge on halo and accuracy (cf. Kozlowski et al., 1986).

When raters evaluated unfamiliar targets, a boundary effect similar to that reported by Kozlowski et al. (1986) was observed that destroyed this clear pattern of results. Raters in the high-job-knowledge condition exhibited an increase in CS–rating and a decrease in rating–true-score covariation relations, although the changes were not significant. High-job-knowledge raters also showed a significant increase in halo and significant decreases in all four accuracy components. These outcomes are consistent with the predictions of the SDH.

Medium- and low-job-knowledge raters evaluating unfamiliar targets, however, showed significant drops in the CS–rating

covariation relation and virtually no change in the rating–true-score covariation association. These findings were also reflected in the halo and accuracy results. Medium- and low-job-knowledge raters exhibited nonsignificant but surprising drops in dimensional intercorrelations and increases in within-ratee SSDs. Medium-job-knowledge raters showed significantly less differential, overall, and correlational accuracy. Stereotypic accuracy was lower, but not significantly so. Low-job-knowledge raters showed lower accuracy overall, although the differences were not significant.

These results for the low-familiarity-ratee condition are not consistent with the SDH, which predicts that relations between CS schemata and rating covariance should increase as familiarity with the ratees decreases, resulting in a weaker rating–true-score covariation association, increased halo, and decreased accuracy. The results were consistent with the SDH only for high-job-knowledge raters, and even then they were in the appropriate direction but not significant. Although this boundary effect was anticipated, the study was not designed to directly evaluate a theoretical model explaining this phenomenon.

One explanation for these results focuses on the amount and quality of prior information available in memory for the medium- and low-job-knowledge groups rating the unfamiliar ratees. For example, Jacobs and Kozlowski (1985) had students rate their college instructors in the 1st, 5th, and 10th weeks of the term. They reported that four operationalizations of halo consistently *increased* with familiarity, the largest increase occurring between the 1st and 5th weeks. They had anticipated that the greatest amount of halo would occur under low-familiarity conditions when raters lacked performance-relevant ratee information. Like the results of the present study, this finding indicates that increased covariation emerged in the ratings when raters possessed some relevant ratee information in memory.

Shweder's (1982) formulation of the SDH specifies that systematic bias in the direction of CS schemata occurs when there is an opportunity for memory decay prior to recall and 20% to 30% error variance is present in ratings. He suggested that raters mistake their implicit semantically based notions of "what goes with what" for actual behavioral covariance under these conditions. Shweder (1980) has been careful to assert that the SDH addresses distortions of covariance and does not deal with issues related to the level of ratings.

What happens when there is no information in memory to recall? A logical extension of Shweder's reasoning would imply that the effect of CS schemata on the patterning of rating covariance would be greater when raters lack specific ratee information in memory. But this might not hold if the operation of the schemata requires *some* recalled information to serve as a level anchor or reference about which the pattern of covariation dictated by the schema is deployed. In the absence of an overall or specific dimensional-level reference, ratings may simply be unpatterned.

Many cognitive heuristics have been implicated in judgment processes (Cantor & Mischel, 1979). In performance-appraisal research, prototypes and relational schemata have been the most prevalent cognitive structures thought to influence rating processes (Nathan & Lord, 1983), although there has been little integration of these two approaches in the literature (cf. Bor-

man, 1983). Prototypes have been conceptualized as abstract or concrete exemplars of categories that provide cognitive economy by allowing a judge to differentiate complex stimuli on the basis of a few salient features (Cantor & Mischel, 1979; Mirvis & Rosch, 1981). Knowledge of the category membership of a stimulus allows a judge to draw inferences about related, although unobserved or unremembered, features (Feldman, 1981).

Categories are thought to be hierarchical in nature, ranging from superordinate, to basic, to subordinate levels of abstraction (Rosch, 1978), with greater differentiation among categories at the superordinate level and greater overlap of category features at the subordinate level (Cantor & Mischel, 1979; Mirvis & Rosch, 1981). In addition, implicit covariance assumptions for features may coexist with the prototypic system at subordinate levels. Nathan and Lord (1983) concluded that dimensional schemata were most strongly implicated in performance ratings, but that a general prototype linked to performance level also affected ratings. Because the target stimuli in this study shared many common features, category assignment would require that raters possess well-developed, concrete, and specific prototype and exemplar systems at the subordinate level. Cantor and Mischel (1979) suggested that experience and interaction with the target domain is necessary for the development of fine-grained subordinate prototypes that reflect the normative characteristics of the stimulus domain. Moreover, Romer and Revelle (1984) suggested that semantic labels guide the encoding and recall of observed behavior. Indeed, they suggested that CS schemata, when used for encoding and recall, should reflect normative behavioral covariation. Thus, raters with greater observational experience and knowledge of a performance domain are likely to have better developed, more concrete, and more specific prototypic systems than do less knowledgeable raters. Moreover, knowledgeable raters are more likely to encode and recall performance-relevant information for specific targets. This implies that job-knowledgeable raters would be better able to assign target ratees to categories (Cantor & Mischel, 1979; Feldman, 1981), thereby providing a level reference. Specific information about the target would be recalled, with the CS schema providing a guide for rating dimensions on which the rater lacked, or could not recall, specific information. Less knowledgeable raters would have poorly developed prototypic systems. Such systems might be sufficient for level assignments when the rater possessed some minimal information about the target ratee, but completely inadequate for category assignments when raters were unfamiliar with ratees. Lacking a level reference, a relational schema such as CS could not by itself provide an adequate guide for ratings. Ratings would probably represent guessed, capricious, and unpatterned responses. We need to explore this boundary and learn what cognitive processes raters use when they lack relevant information for making ratings. Research that takes a process-tracing (Payne, Braunstein, & Carroll, 1978) approach may help address this question.

The halo results were surprising, as individuals in all conditions consistently underestimated the true amount of dimensional intercorrelation. One explanation for this phenomenon, that the level of intercorrelation among the true scores exceeded the estimates typically provided in other studies, proved to be

untenable. We examined the literature and found seven studies that reported the level of intercorrelation among objective performance dimensions (Alexander & Wilkins, 1982; Anderson, Roush, & McClary, 1973; Bass & Turner, 1973; Borman, 1979; Cascio, & Valenzi, 1978; Kirchner, 1960; Seashore, Indik, & Georgopoulos, 1960). The level of intercorrelation ranged from $-.02$ to $.87$, with a mean level of $.39$. Values for the present study were quite similar. High-familiarity-ratee intercorrelations ranged from $.02$ to $.93$, with a mean of $.40$, whereas low-familiarity-ratee intercorrelations ranged from $.00$ to $.96$, with a mean of $.45$.

An alternative explanation is that the specificity of the performance dimension labels used in the present study made the dimensions appear to be more distinctive to the raters than they actually were. Yet, because only a relatively small set of physical abilities underlie these dimensions (Gould, 1985) and the dimensions are not independent, they are moderately to highly intercorrelated.

However, the most parsimonious interpretation may well be that halo error is not nearly as serious a problem as has generally been presumed. The determination of halo error has generally been on the basis of observed rating intercorrelations that exceed researchers' expectations of the level of covariation among dimensions regarded as conceptually distinct. Studies that purport to demonstrate halo error via average intercorrelations or rater standard deviations without reference to true scores that are independent of human judgments, are unanchored. Without knowledge of the levels of true-score intercorrelation, it is difficult to maintain that halo is a pervasive error. The fact is, we neither know how pervasive it is nor do we clearly understand what its effects on accuracy are.

The ANOVA results for rating accuracy were consistent with conventional views; greater job and ratee knowledge lead to greater accuracy. The correlational results for accuracy were more intriguing. Conceptual similarity-rating covariation was positively associated with differential accuracy and negatively associated with stereotypic accuracy. The same relations were observed for halo and accuracy. These relations were consistent across both ratee-familiarity conditions. Although research has seldom directly examined these relations, studies that have directly examined them have generally reported positive associations between covariation and differential accuracy (Murphy & Balzer, 1986). This finding strongly implicates relational schemata or prototypes, or both, in rating processes that result in accurate evaluations.

Limitations

Addressing the SDH at the appropriate individual level rectified a serious methodological problem in prior research, but it also introduced new problems. In all, 69 cases had to be dropped from the analysis because individual rating intercorrelation matrices could not be computed. This is not an issue that arises when data are aggregated over raters. Rating all ratees with the same value for a dimension is a legitimate rating response, albeit often regarded as an error, but it could not be examined at the individual level. We don't believe that this problem significantly affected the results (see Footnote 4). Yet, it is exactly this tendency, lack of dimensional variance, that

results in greater halo and closer CS-rating covariation. We do not know of a way around this problem, and it is a potentially serious one for further rating-process studies at the individual level.

Because the covariation, halo, and accuracy indices were computed within raters, the number of ratees determined the stability of the individual-level measures. Unless researchers are able to have raters make a substantial number of ratings, such individual-level measures will always be based on a small sample. Even the present study with its modest sample required the raters to make 276 judgments. Increasing the number of ratings to be made, however, puts tremendous demands on both raters and researchers. This problem suggests that there are practical limits to our ability to investigate relational schemata at the individual level.

There were also a number of issues that may limit the generalizability of the results reported here.⁵ First, it is always difficult to generalize results derived from student raters to real appraisal situations. For example, it is unlikely that the sample included raters at the very top of the range of job knowledge in the domain of baseball performance. This range restriction on expertise is far less likely in organizations. Similarly, rater observations of ratee performance were conducted under incidental learning conditions, whereas in organizations at least some of a manager's time is spent on active appraisal. Although it is unlikely that either of these issues affects our conclusions, they do limit the direct comparability of the results to organizational settings.

Second, the performance domain and the dimensions used in the present study may represent a somewhat unusual case in performance-appraisal research. For example, performance on the baseball dimensions depends on a relatively small set of primarily physical abilities (Gould, 1985), whereas the more usual appraisal situation focuses on dimensions representing a broader mix of cognitive and dispositional factors. Moreover, baseball players are selected to be high on several dimensions (cf. Cooper, 1981a, 1981b; Gould, 1985), producing low interdimensional variance. Indeed, there may be definite limits on just how much baseball performance can vary (Gould, 1985). It is possible that stronger results would be obtained for ratee targets who are more variable in their performance.

Research Implications

The results of the present study and our effort to interpret them have raised several questions regarding the operation of relational schemata in judgment processes. The investigation of the SDH within the performance appraisal domain has yielded some support (Kozlowski et al., 1986), but has also shown the effect to be far less robust than was reported by Shweder. Across seven studies, Shweder (1982) reported intermatrix correspondence relations of .75 for CS-rating covariance and .25 for rating-true-score covariance. The differences in magnitude between these values are much larger than comparable differences shown in Table 1. This difference in results is partly due to the group level of data and analysis used by Shweder, in comparison to the theoretically appropriate individual level used in the present study (L. James, 1982). The disparity may also be partly due to differences in the level of abstraction between the trait

labels used in Shweder's research and our use of more concrete behavioral frequency labels. Abstract trait labels are more likely to foster ambiguity and increased covariation among rated dimensions than are specific behavior labels. More important, Shweder's results are also due to the use of biased true scores that underestimate true behavioral covariation and thus make it appear as if rating covariation is insensitive to actual covariation (Romer & Revelle, 1984).

The results of the present study and previous research (Kozlowski et al., 1986) indicate that implicit covariance schemata play a role in the rating process and affect halo and accuracy. They also indicate that a reformulation of the SDH to account for the observed boundary effect is warranted. One problem to doing so is that the SDH is underspecified with respect to a theory of cognitive structure and processing and, in addition, makes several assumptions that have escaped close scrutiny. We previously outlined a more detailed framework to explain the operation of relational schemata in ratings. One hypothesis suggested that job-knowledgeable raters have better developed prototypic systems for the relevant job domain than do less knowledgeable raters. This could be examined by contrasting the number of categories, number and kind of features, and the vividness and concreteness of features generated by raters with different levels of experience within a given job domain (cf. Cantor & Mischel, 1979).

Shweder (1975, 1982) assumed that CS schemata are semantically driven and culturally normative, although erroneous. These assumptions imply that there should be high interrater agreement on CS schemata, that CS schemata should be reliable over time, that CS schemata should not be strongly related to true-score covariation, and that individual differences in job knowledge should have little effect on the structure of CS relations. The framework presented previously asserts that there should be differences in the structure of CS schemata by job knowledge. The schemata of high-knowledge raters should be more reflective of normative behavioral covariation (Romer & Revelle, 1984). Although the present study examined the equivalence of the average level of CS relations, it did not address the issue of different CS schemata structures. The study also did not examine CS-true-score covariation relations. Whether CS schemata are culturally based or derived from true behavioral covariance, the sampling error present in a true-score covariance matrix based on a sample of 10 is too great to draw valid conclusions. This question is best addressed by examining the structure of CS schemata with multidimensional scaling and relating the structure of CS relations to true-score covariance derived from a large sample.

Finally, the issue of the boundary condition warrants closer inspection. Something different happened in the rating process when raters had little or no knowledge about the targets they evaluated. The responses appeared to be unpatterned and, in and of themselves, may be of little interest. However, the minimum conditions necessary to bring structure to the ratings, in terms of covariation relations, halo, and accuracy, have the potential to help illuminate the cognitive processes that operate

⁵ Many of the limits to the generalizability of this study were suggested by an anonymous reviewer.

when raters have information available in memory. There have been many calls for more process-oriented research, and there is certainly no shortage of theoretical models. With few exceptions (e.g., Feldman, 1981), most of these models are underspecified with respect to raters' cognitive operations. Perhaps what is needed is some qualitative, descriptive research to flesh out the processes per se. Process-tracing research (e.g., Payne et al., 1978) that focuses on the rater's amount, search for, and use of performance-relevant information may yield a better understanding of rater-judgment processes and provide a basis for theoretical development in this area.

References

- Alexander, E. R., & Wilkins, R. D. (1982). Performance rating validity: The relationship of objective and subjective measures of performance. *Group and Organization Studies*, 7, 485-496.
- Anderson, H. E., Roush, S. L., & McClary, J. E. (1973). Relationships among ratings, production, efficiency, and the general aptitude test battery scales in an industrial setting. *Journal of Applied Psychology*, 58, 77-82.
- Bass, A. R., & Turner, J. N. (1973). Ethnic group differences in relationships among criteria of job performance. *Journal of Applied Psychology*, 57, 101-109.
- Bingham, W. V. (1939). Halo, invalid and valid. *Journal of Applied Psychology*, 23, 221-223.
- Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance*, 20, 238-252.
- Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology*, 57, 15-22.
- Borman, W. C. (1983). Implications of personality theory and research for the rating of work performance in organizations. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance measurement and theory* (pp. 127-172). Hillsdale, NJ: Erlbaum.
- Cantor, N., & Mischel, W. (1979). Prototypes in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 12, pp. 3-52). New York: Academic Press.
- Cascio, W. F., & Valenzi, E. R. (1978). Relations among criteria of police performance. *Journal of Applied Psychology*, 63, 22-28.
- Cooper, W. H. (1981a). Conceptual similarity as a source of illusory halo in job performance ratings. *Journal of Applied Psychology*, 66, 302-307.
- Cooper, W. H. (1981b). Ubiquitous halo. *Psychological Bulletin*, 90, 218-244.
- Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." *Psychological Bulletin*, 52, 177-193.
- D'Andrade, R. G. (1974). Memory and the assessment of behavior. In T. Blalock (Ed.), *Measurement in the social sciences* (pp. 159-186). Chicago: Aldine.
- Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology*, 66, 127-148.
- Gould, S. J. (1985). *The flamingo's smile: Reflections in natural history*. New York: Norton.
- Jacobs, R., & Kozlowski, S. W. J. (1985). A closer look at halo error in performance ratings. *Academy of Management Journal*, 28, 201-212.
- James, B. (1982). *The Bill James baseball abstract, 1982*. New York: Ballantine.
- James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology*, 67, 219-229.
- Kirchner, W. K. (1960). Predicting ratings of sales success with objective performance information. *Journal of Applied Psychology*, 44, 398-403.
- Kozlowski, S. W. J., Kirsch, M. P., & Chao, G. T. (1986). Job knowledge, ratee familiarity, conceptual similarity, and halo error: An exploration. *Journal of Applied Psychology*, 71, 45-49.
- Mirvis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, 32, 89-116.
- Murphy, K. R., & Balzer, W. K. (1986). Systematic distortions in memory-based behavior ratings and performance evaluations: Consequences for rating accuracy. *Journal of Applied Psychology*, 71, 39-44.
- Nathan, B. R., & Lord, R. G. (1983). Cognitive categorization and dimensional schemata: A process approach to the study of halo in performance ratings. *Journal of Applied Psychology*, 68, 102-114.
- Newcomb, T. (1931). An experiment designed to test the validity of a rating technique. *Journal of Educational Psychology*, 22, 279-288.
- Payne, J. W., Braunstein, M. L., & Carroll, J. S. (1978). Exploring predecisional behavior: An alternative approach to decision research. *Organizational Behavior and Human Performance*, 22, 17-44.
- Pulakos, E., Schmitt, N., & Ostroff, C. (1986). A warning about the use of a standard deviation across dimensions within ratees to measure halo. *Journal of Applied Psychology*, 71, 29-32.
- Romer, D., & Revelle, W. (1984). Personality traits: Fact or fiction? A critique of the Shweder and D'Andrade systematic distortion hypothesis. *Journal of Personality and Social Psychology*, 47, 1028-1042.
- Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 28-49). Hillsdale, NJ: Erlbaum.
- Seashore, S. E., Indik, B. P., & Georgopoulos, B. S. (1960). Relationships among criteria of job performance. *Journal of Applied Psychology*, 44, 195-202.
- Shweder, R. A. (1975). How relevant is an individual difference theory of personality? *Journal of Personality*, 43, 455-484.
- Shweder, R. A. (1980). Factors and fictions in person perception: A reply to Lamiell, Foss, & Cavenee. *Journal of Personality*, 48, 75-81.
- Shweder, R. A. (1982). Fact and artifact in trait perception: The systematic distortion hypothesis. In B. A. Maher & W. B. Maher (Eds.), *Progress in experimental personality research: Normal personality processes* (Vol. 11, pp. 65-100). New York: Academic Press.
- Shweder, R. A., & D'Andrade, R. G. (1980). The systematic distortion hypothesis. In R. A. Shweder & D. W. Fiske (Eds.), *New directions for methodology of behavioral science: Fallible judgment in behavioral research* (pp. 37-58). San Francisco: Jossey-Bass.
- Wiggins, J. (1973). *Personality and prediction: Principles of personality assessment*. Reading, MA: Addison-Wesley.

Received September 26, 1986

Accepted October 29, 1986 ■

Further Investigation of Common Knowledge Effects on Job Analysis Ratings

Angelo S. DeNisi and Edwin T. Cornelius III
College of Business Administration, University of South Carolina

Allyn G. Blencoe
Blue Cross/Blue Shield of South Carolina

Previous research (Smith & Hakel, 1979) raised the possibility that the Position Analysis Questionnaire (PAQ) only captures common knowledge, or stereotypes, about jobs. Cornelius, DeNisi, and Blencoe (1984) presented data to refute this, but found that the number of PAQ items rated *does not apply* (DNA) was related to the agreement between naive raters and expert raters. The present study used data from 87 analysts and 24 jobs. Naive ratings were those obtained from analysts who had not studied the job, whereas expert ratings were those obtained from raters who had observed the job, interviewed incumbents, and written task statements describing the job. These ratings were then compared to target-score profiles obtained from PAQ services. Results confirmed earlier hypotheses that large numbers of DNA items artifactually inflate correlational estimates of agreement between expert and naive raters. In addition to this artifact, results also supported the view that the PAQ is less appropriate as a job analysis tool for some types of jobs. Implications for research and practice are discussed.

Recently, questions have been raised about the validity of ratings obtained on structured job analysis instruments. The validity of these ratings has generally been taken for granted; however, given the range of applications for these ratings, these questions represent potentially serious problems for human resource practitioners and researchers.

Most of the attention in this area has focused on the Position Analysis Questionnaire (PAQ; McCormick, Jeanneret, & Mecham, 1972). The PAQ is the most frequently used job analysis instrument in the published literature (Cornelius, 1987), and so it follows that there should be a great deal of concern about the validity of the data generated by the PAQ. Furthermore, PAQ ratings have been used in a variety of applications, such as those concerned with establishing wage and salary levels for jobs (Robinson, Wahlstrom, & Mecham, 1974), identifying aptitude tests that can be used as valid selection devices (McCormick, DeNisi, & Shaw, 1977), and establishing job classification systems (Taylor, 1978).

Questions about the validity of PAQ ratings began with a study published by Smith and Hakel (1979) in which the authors reported extremely high correlations (in the mid 90s) between PAQ ratings obtained from expert raters (analysts, incumbents, and supervisors) and those obtained from naive rat-

ers (college students) relying only on job titles or on job titles plus brief job descriptions. The implication is that if students who have never studied a job can provide ratings that are highly correlated with those provided by experts, the PAQ may reflect only common knowledge about jobs. Stated differently, these results raise the possibility that the PAQ is capable only of measuring job stereotypes. Such a possibility is extremely troublesome for advocates of the PAQ, as well as for advocates of other structured job analysis instruments.

The seriousness of this potential problem motivated subsequent studies. Cornelius, DeNisi, and Blencoe (1984), using the same paradigm as Smith and Hakel (1979), used a more appropriate method for calculating the convergence between expert and naive raters. This method produced correlations in the fifties instead of in the nineties, and in general, the data suggested that the PAQ was able to provide job information beyond what would be considered common knowledge or stereotypes.

Although these results suggest there is no reason for panic, some of the findings reported by Cornelius et al. (1984) are reason for concern. These findings describe the impact of PAQ items that are rated *does not apply* (DNA) by the analyst. Because the PAQ was designed to be used with virtually any job, some items are always rated DNA for any particular job. But it can be shown that the number of DNA items has a confounding effect on correlations calculated between PAQ profiles (Harvey & Hayes, 1986). It is also possible to view the number of DNA items as an index of the appropriateness of the PAQ for analyzing a particular job. That is, when large numbers of PAQ items are found not to apply, the level of information about the job available from the PAQ may be too limited to provide a meaningful description of the job. In such cases, naive analysts using the PAQ may be able to provide data comparable to that of expert raters. This second possibility is the focus of the present

Some of the data in this article were reported as part of a symposium at the meeting of the American Psychological Association, Los Angeles, 1985. Comments by the discussant, Walt Tornow, were useful in preparing the present version. We are also indebted to the editor and three reviewers for helpful comments on an earlier version of the article.

Correspondence concerning this article should be addressed to Angelo S. DeNisi, Riegel and Emory Human Resource Research Center, College of Business Administration, University of South Carolina, Columbia, South Carolina 29208.

study. Before discussing it further, however, we will review the role of DNA items as methodological artifact.

Does-Not-Apply Items as Methodological Artifact

Each item on the PAQ includes a response category indicating that the individual item does not apply to the job in question. As pointed out by Cornelius et al. (1984), the correlation between naive raters and expert raters (the paradigm for establishing the impact of common knowledge on PAQ responses used by Smith & Hakel, 1979) may be artifactually high due to the large numbers of PAQ items rated DNA for a particular job. For example, most people could correctly determine that the job of college professor does not require the use of "long-handled tools" (an item on the PAQ), even without carefully studying the job. Thus, naive raters would agree perfectly with expert raters in their ratings for this item—the item "does not apply." There would, of course, also be a number of other items that obviously do not apply, and raters would also be in agreement in rating these items. But the agreement on these ratings would increase the correlation between expert and naive raters, even if they generally disagreed on their ratings for the PAQ items that did apply to the job. As we noted, it is this correlation that has been interpreted (in part) as evidence that the PAQ can only provide descriptions of common knowledge about jobs.

Clearly then, the number of PAQ items rated DNA for a job can distort the true extent to which expert and naive ratings agree. In fact, Harvey and Hayes (1986) provided graphic evidence of the extent of this distortion. They used Monte Carlo techniques to vary the number of DNA items systematically. They found that if 21% of the items were rated DNA for a job, the correlation between two PAQ profiles could be as high as .50, *even though they used completely random data* for the non-DNA items. Likewise, if 56% of the items were rated DNA, the correlation for random data was .78. These results provide a baseline for interpreting naive-expert PAQ ratings and help explain why Cornelius et al. (1984) found that for the job of faculty member (where 46% of the items were rated DNA), eliminating the DNA items dropped the correlation between naive and expert raters from .77 to .43.

Does-Not-Apply Items as a Measure of Appropriateness

As we noted, when DNA items were omitted, Cornelius et al. (1984) found that correlations between expert and naive ratings dropped from .77 to .43. However, these authors also found that for the job of maintenance foreman (where only 15% of the items were rated DNA), omitting the DNA items had no effect on the correlation between expert and naive raters ($r = .49$, with DNA items, and $r = .47$, without). Although this pattern of results follows from the relation reported by Harvey and Hayes (1986), it raises an interesting question: What does it mean when almost one half the items on the PAQ are rated DNA? One possibility is that the job demands little of the incumbent in either mental or physical activity (i.e., the job is too simple). Another possibility is that the nature of those demands is so different from those commonly found that few, if any, job analysis procedures should be able to detect them. The PAQ, due to its generic nature (i.e., it is appropriate for use with a wide range

of jobs), might be a particularly insensitive device in this situation. If the PAQ is used for such a job, we should not be surprised to find that naive analysts can provide information that is correlated with that provided by expert raters. In fact, it is possible that in such cases, the PAQ can provide little information beyond commonly held stereotypes about the job.

Note that significant correlations between expert and naive ratings can be the result of either a methodological artifact or the inappropriate use of the PAQ, or both. The effect of DNA items as a methodological artifact can be adjusted for mathematically. This could be accomplished by recomputing correlations deleting DNA items, or by some other transformation of the data to minimize the effects of the DNA on any correlations. The "appropriateness" problem, on the other hand, cannot be made to disappear so easily, for this implies that the PAQ should not have been used to analyze the job in the first place. One purpose of the present study, then, was to test the working hypothesis that the number of PAQ items rated DNA can be seen as an index of the appropriateness of the PAQ for analyzing a particular job. This role is distinct from the role of DNA items as a methodological artifact, and in this sense the number of DNA items may be used to help identify jobs for which expert and naive job analysis ratings may well be equivalent.

Specifically, we designed the present study to extend earlier work, especially the Cornelius et al. (1984) study, in several areas. One of the shortcomings of the earlier studies is that expert and naive ratings could only be compared with each other. Given the relatively small number of expert raters per job, it was possible that even the expert ratings did not accurately describe the jobs in question, due to measurement error. In the present study we attempted to deal with this problem by using data from PAQ Services, Logan, Utah (which scores all PAQs and maintains a file of more than 60,000 individual analyses) to form target-score profiles for each job, which we could then use as a basis for a more meaningful comparison of expert and naive ratings. In addition, the Cornelius et al. (1984) study was not designed to test the role of DNA items in determining the convergence between expert and naive raters, but rather was designed to determine if the Smith and Hakel (1979) results would be replicated using alternative computational procedures. Therefore, there was no attempt to study jobs for which greater or lesser numbers of PAQ items applied. In the present study, jobs were studied for which the number of DNA items was expected to vary considerably. This made it possible to examine the role of DNA items as both a computational artifact and an index of PAQ appropriateness. Finally, we attempted to determine if there were some way of predicting jobs for which PAQ items were more or less applicable. Specifically, because many of the PAQ items concern the use of tools or equipment (more than 35 items make direct reference to using tools on the job), we selected jobs that varied in terms of scores on the Things scale from Functional Job Analysis (Fine & Wiley, 1971). Scores on this scale indicate the extent to which a job involves the use of tools, machines, and equipment.

As a framework for extending this past research, we tested the following hypotheses:

Naive PAQ ratings will be less reliable than expert PAQ ratings.

The correlations between expert and naive PAQ ratings will be significantly different from 1.00, indicating that the two sets of ratings cannot be considered equivalent. These first two hypotheses represent a replication of the Cornelius et al. (1984) findings, with expert ratings defined as those obtained after a rater studied the job in question. Although the second hypothesis is stated in terms of the null, it too reflects the results reported by Cornelius et al. (1984). Specifically, it addressed the psychometric requirement that two sets of scores (ratings) that are equivalent must be perfectly correlated (Gulliksen, 1968). In addition, because the earlier authors found that the convergence between expert and naive raters was reduced when items rated DNA were omitted, the following hypothesis was tested as the logical extension of their results:

The correlation (convergence) between expert and naive raters will vary directly as a function of the number of items rated DNA. That is, the correlation between expert and naive raters will be higher for jobs for which fewer items apply. This is a direct test of the role of DNA items as a computational artifact, following the suggestions of Harvey and Hayes (1986). Furthermore, the availability of averaged PAQ profiles from the PAQ data base made it possible to correlate each set of ratings with target profiles that are less susceptible to individual rater idiosyncrasies. These correlations are referred to as fidelity coefficients, inasmuch as they reflect the extent to which each set of ratings faithfully reproduced the pattern of ratings in the target profile. Consistent with the implications of the Cornelius et al. (1984) study, we therefore hypothesized that

Fidelity coefficients for expert ratings will be significantly higher than those for naive ratings. The fifth hypothesis addresses the distinction between the two proposed roles for DNA items. If DNA items can be viewed as an index of the appropriateness of the PAQ, there should be a relation between the number of DNA items and the correlation between PAQ ratings and the PAQ Services' target-score profiles, *but only for naive raters.* That is, as the number of DNA items increases and the sensitivity of the PAQ for the job decreases, we would expect naive analysts to provide information that is as good as that provided by the PAQ data base. Therefore, fidelity coefficients for naive analysts should be lower when there are fewer DNA items and higher when there are many DNA items. In other words, we expect a high correlation between the number of DNA items and the naive rater fidelity coefficients. In addition, this correlation should be significantly weaker for the expert analyst ratings because their ratings are expected to vary as a function of differences in actual job characteristics rather than differences in the number of items rated DNA. Notice that this hypothesis is crucial to understanding the impact of DNA items on the validity of PAQ data. If DNA items function only as a methodological artifact (and can therefore be corrected mathematically), we would expect no differences in the correlation between fidelity coefficients and the number of DNA items for expert and naive analysts. Thus, our fifth hypothesis:

Naive fidelity coefficients will be significantly correlated with the number of items rated DNA, and this correlation will be significantly greater than that for expert fidelity coefficients. Note that evidence for this hypothesis would also provide additional support for the idea that expert analysts can provide informa-

tion beyond that which is available from naive analysts—at least for jobs where the PAQ is appropriate.

The sixth and final hypothesis is largely exploratory and concerns a systematic approach to determining jobs that are appropriate for testing correlational hypotheses using the PAQ, as indicated by the number of items rated DNA. Specifically,

There will be fewer items rated DNA for jobs scoring higher on the Things scale from Functional Job Analysis. As we noted earlier, because a considerable number of PAQ items concern the use of tools, equipment, and machinery, the Things scale (which assesses involvement with equipment and machinery) would seem to be a good indicator of those jobs for which large numbers of PAQ items will not apply. The number of DNA items, again, is viewed as an index of the appropriateness of using the PAQ for a given job. Note that low scores on the Things scale indicate higher levels of involvement.

Method

Subjects and Procedures

Subjects were 87 university students enrolled in two sections of an undergraduate personnel management course. As part of the course requirement, students, in teams of 3 to 5 members, completed a job analysis project. They were required to read job materials, observe incumbents performing the job, and interview at least three incumbents or supervisors. Students were also required to write task statements according to the format specified in Fine and Wiley (1973).

Prior to beginning the project, all of the students were given training in job analysis methods. This training consisted of one class lecture on general job analysis methods and three lectures on task-oriented job analyses, including how to interview job incumbents and how to write task statements. Students were also assigned outside reading in the Department of Labor's (1972) *Handbook for Analyzing Jobs* and Fine and Wiley's (1971) *An Introduction to Functional Job Analysis*. As an incentive for conducting a thorough analysis, performance on this project received substantial weight (20%) in the final course grade. To increase student motivation, the instructor also conducted a review session with each team midway through the semester. During this session, preliminary task statements were reviewed and critiqued, suggestions were made, and questions were answered.

Sample of Jobs

Each student team submitted for approval the titles of one or more jobs they wished to study. Teams were instructed to choose jobs for which they would have easy access to incumbents and supervisors. The instructor approved jobs for analysis so that (a) there were no redundancies across teams and (b) there was a reasonable distribution of jobs on the Things scale. Following this procedure, 24 different jobs were approved for analysis. These jobs and their corresponding *Dictionary of Occupational Titles* (DOT; Department of Labor, 1977) codes are listed in Table 1.

Naive and Expert Ratings

Naive and expert PAQ ratings were obtained on each of the 24 jobs. The naive ratings were collected at the beginning of the semester by having students fill out the PAQ, given only a job title. An effort was made to ensure that students had little or no knowledge about the jobs they were evaluating. This was done by having students rate their familiarity with each of the 24 jobs using a 5-point scale that ranged from *I have no idea what is done on this job* (1) to *I have done this job myself*

or know someone well who has done it (5). The instructor then picked students who were *not* familiar with the job in question to provide PAQ ratings (i.e., rated the job either 1 or 2 on the familiarity scale). In this way, each of the jobs in this study was independently rated by from 3 to 5 students who had no intimate information about the job and who thus could be considered *naive*.

Expert PAQ ratings were also obtained on each job. At the end of the semester, one class period was devoted to students filling out the PAQ for the job they had analyzed during the semester. These ratings were considered *expert* because at this point the students were highly familiar with the jobs. It should be noted that no one analyzed the same job as both a naive and an expert rater.

Target-Score Ratings

To evaluate the fidelity of the naive and expert ratings, we obtained a profile of PAQ target-score ratings for each job. Because PAQ Services had PAQ data for more than 60,000 analyses, it was possible to obtain an average PAQ profile that was based, in some cases, on hundreds of individual analyses. Inasmuch as the average profiles from the Utah data base are less subject to the idiosyncrasies inherent in any single position, these data thus provided reasonable target scores for the jobs in question.

For each job analyzed by the students, the authors obtained a six-digit code number from the *Dictionary of Occupational Titles* (Department of Labor, 1977). These codes were then transmitted to PAQ Services to determine the jobs for which average profiles were available. An arbitrary decision rule was established that a PAQ profile must be based on at least 50 individual analyses before it could be used in this study. Using this decision rule, target-score profiles were available for 13 of the 24 jobs in the sample.

Number of Does-Not-Apply Items for Each Job

The number of PAQ items that did not apply was calculated using PAQ responses from the expert sample (i.e., students who had studied the jobs all semester). For those student teams consisting of 5 members, 4 of the 5 raters must have indicated that the item did not apply before the item was counted as inappropriate. For those student teams consisting of 3 or 4 members, all of the raters must have indicated that the item did not apply. These criteria are similar to those applied in the Cornelius et al. (1984) study.

Involvement With Things Scale

Because it was hypothesized that there would be fewer DNA items for jobs with higher levels of involvement with machines and equipment, a quantitative measure of this involvement was needed. The final digit of the six-digit DOT code represents a job's score on the Things scale from Functional Job Analysis. These scores can range from a low of 7 (feeding-offbearing) to a high of 1 (precision working). For the present study, jobs were classified as being high on the Things scale (scores of 1 or 2), moderate (scores of 3 or 4), or low (all remaining jobs in the sample had a score of 7).

Data Transformation

One final point needs to be made concerning the way in which data were analyzed. As we noted earlier, the Cornelius et al. (1984) study reported that for some jobs almost half of the items from the PAQ may be rated DNA. In addition, Harvey and Hayes (1986) reported that the average number of DNA items was 100 ($SD = 21$), or 51% of all items, for the 90 different municipal jobs they studied. Thus, it is conceivable that for any given job, more items receive a rating of DNA (0) than any

other rating (most PAQ items use 5-point scales, not counting the DNA category), producing a skewed distribution of item responses. To make correlations computed with such distributions more meaningful and to reduce the effect of the DNA items on any such correlations, all of the data in the study were first transformed using a log transformation before any other analyses were conducted.¹ Thus, all of the results reported are based on these transformed data.

Results

The results from the study are summarized in Table 1. The 24 jobs studied are listed with their complete six-digit DOT codes (the final digit represents the score on the Things scale). The third column presents the number of items rated DNA for each job, using the procedures described earlier. Interrater reliability estimates are then presented for both the naive and expert raters. These entries represent the mean of all possible pairwise correlations among raters—corrected for the number of raters using the Spearman-Brown formula—and were computed in the manner recommended in Nunnally (1978).² Also presented are the mean correlations between expert and naive raters for each job. These correlations were computed by taking the average item rating for naive raters on each of the 187 PAQ items and the average item rating for expert raters on each item, and correlating these averages across the 187 items. These same average item ratings were used to compute the naive and expert fidelity coefficients presented in the last two columns of Table 1. In each case, the average item ratings were correlated with the item ratings from the target-score profiles obtained from PAQ Services.

In the first hypothesis we predicted that expert raters would be more reliable (i.e., greater agreement among raters) than would naive raters. As can be seen in Table 1, this was the case for all but two jobs (spa health instructor and insurance salesman). Mean reliability estimates were calculated for expert and naive raters (after r to z transformations) and then compared. The mean expert reliability (.85) was found to be higher than the mean naive reliability (.72), although the difference failed to reach the conventional .05 significance level ($z = 1.60$, $p < .06$, one-tailed). The results thus provide only modest support for the first hypothesis.

The second hypothesis concerned the correlations between expert and naive ratings for each job. As shown in Table 1, these correlations ranged from .54 to .91, with a mean of .72. Each correlation was converted to a z value and tested to see if it differed significantly from 1.00. All except the correlation for personnel manager (.91) differed from 1.00 ($p < .05$), providing support for the second hypothesis.

The third hypothesis we stated predicted that naive–expert correlations would vary as a function of the number of items rated DNA. The number of DNA items was correlated with the expert–naive correlations (transformed using the r to z transformation), resulting in a correlation of .50 ($p < .01$). Thus, there was support for the third hypothesis.

¹ We are grateful to the editor for suggesting this analytic approach.

² These correlations were not corrected for rater unreliability, however, inasmuch as the lack of reliability—at least on the part of naive raters—is extremely relevant to the arguments made in this article.

Table 1
Reliability and Fidelity Data for Jobs Studied (After Log₁₀ Transformation)

DOT code ^a	Job title	No. of DNA items	Reliability		Expert- Naive ^b	Fidelity	
			Naive	Expert		Naive	Expert
332.271	Hairdresser	21	.63	.68	.59	—	—
824.261	Electrician	16	.69	.79	.68	.32	.49
723.381	Appliance service representative	5	.68	.79	.66	.28	.40
079.361	Respiratory therapist	15	.72	.79	.72	—	—
720.281	TV repairman	12	.60	.70	.68	—	—
199.361	X-ray technician	16	.70	.81	.57	.29	.49
078.362	EKG technician	27	.79	.80	.66	—	—
292.353	Beer truck driver	22	.65	.81	.56	—	—
375.263	Police officer	19	.53	.81	.77	.49	.57
906.683	UPS truck driver	17	.81	.82	.75	.43	.39
312.474	Bartender	35	.67	.84	.63	.44	.41
152.224	Athletic director	10	.76	.88	.76	—	—
827.464	Air conditioning installer	25	.46	.88	.72	—	—
406.687	Groundskeeper	33	.68	.76	.76	.46	.49
869.664	Construction worker	21	.64	.88	.59	.57	.38
099.227	Adult education instructor	62	.77	.85	.79	—	—
166.167	Personnel manager	63	.87	.90	.91	.58	.57
090.227	College professor	50	.81	.97	.76	.63	.59
153.227	Spa health instructor	28	.82	.79	.78	—	—
290.447	Sales clerk	56	.65	.93	.70	.49	.40
250.257	Insurance salesman	32	.89	.88	.84	.47	.49
241.214	Claims adjuster	19	.81	.82	.74	—	—
163.167	Sales manager	21	.80	.83	.78	.48	.45
239.357	Radio time salesman	25	.66	.84	.54	—	—

Note. DOT = Dictionary of Occupational Terms; DNA = does not apply; EKG = electrocardiogram.

^a The sixth digit represents the score on the Things scale from Functional Job Analysis for the job.

^b These are the correlations between average naive and average expert item ratings.

In the fourth hypothesis we predicted that expert fidelity coefficients would be significantly higher than naive fidelity coefficients. The average expert coefficient was .47, whereas the average naive coefficient was .46. Thus, there was no support for the fourth hypothesis.

In the fifth hypothesis we addressed the role of the DNA items as an index of the appropriateness of the PAQ for analyzing a given job, and predicted a positive relation between the number of DNA items for a job and the naive fidelity coefficients, but not for the expert fidelity coefficients. The correlation between the number of DNA items and the naive fidelity coefficients was computed to be .70 ($p < .01$), whereas the correlation with expert fidelity coefficients was computed to be .38 (*ns*). The two coefficients were found to be marginally different from each other ($z = 1.50, p < .07$, one-tailed), providing moderate support for this hypothesis.³

The final hypothesis was exploratory in nature. We predicted fewer DNA items for jobs scoring high on the Things scale (7 is the lowest score, 1 the highest). The mean number of DNA items for jobs scoring lowest on Things (7) was 40; the mean for jobs with moderate scores (3 or 4) was 23; and the mean for jobs scoring highest on Things (1) was 16. These means were significantly different from each other, $F(2, 23) = 4.16, p < .05$, with only the extremes differing from each other. The correlation between the number of DNA items and the Things scores was .66 ($p < .01$), providing strong support for the final hypothesis.

Discussion

The results of the present study shed new light on the question of whether naive raters can provide PAQ ratings equivalent to those of expert raters. There is now further evidence that naive raters generally cannot provide equivalent ratings. However, for jobs where large numbers of PAQ items do not apply, these results suggest that it may be less appropriate to use the PAQ and that neither naive nor expert raters can provide very useful information. Furthermore, there is reason to believe that jobs that score low (3 or higher) on the Things scale may be those for which the PAQ is less appropriate, and that another type of job analysis approach (such as a task-oriented method) may be called for.

Earlier research (Cornelius et al, 1984; Smith & Hakel, 1979) explored the possible equivalence of PAQ ratings from different sources. These studies, along with some other work (Harvey & Hayes, 1986), led to the suggestion that PAQ items rated DNA were important to any comparisons among ratings. But the studies that examined the role of DNA items tended to concentrate on their role as a methodological artifact only. The present study expanded that view by examining the role of DNA items as an index of the appropriateness of using the PAQ for a given

³ There is, of course, a problem of power in computing this difference. In fact, given the same magnitude of difference and only two more cases, the difference would have been significant at the .05 level.

job. Specifically, large numbers of DNA items could indicate that work performed on a job was too simple or was beyond the scope of activities described by the PAQ. In such cases, correlations among ratings from different sources might well be high, but these cases were simply not appropriate for drawing conclusions about the equivalence of different rating sources. This role was supported by the significant correlation between the number of DNA items and the degree of convergence between expert and naive ratings ($r = .50$), as well as by the different relation between DNA items and fidelity coefficients for expert and naive ratings. These coefficients reflected the correlation between each set of ratings and target-score profiles for a set of jobs obtained from PAQ Services, and these coefficients for naive raters were more sensitive to the number of DNA items than were those for expert raters. In fact, the number of DNA items shared 40% of the variance with these fidelity coefficients for naive raters, versus 14% for the expert fidelity coefficients.

But how would researchers or practitioners know a priori which jobs are more or less appropriate for making comparisons among sources of ratings? Because the number of DNA items appears critical to this decision, and a fair proportion of the PAQ items deal with tools and machines, the present study also investigated whether the Things score from Functional Job Analysis (the final digit of the six-digit DOT code) predicted the number of DNA items. Although the present results are far from conclusive, they do suggest that the Things score is a good indicator of the number of DNA items, and thus of the appropriateness of using the PAQ to analyze a given job.

Although the results of the present study indicate that expert and naive PAQ ratings are probably not equivalent, they nonetheless raise a number of questions for future research. One of these questions stems from the fact that almost every study in this area has used the Position Analysis Questionnaire. Yet, exactly because the PAQ is designed to describe work activities on a wide variety of jobs it is prone to the DNA problem we have addressed in this study. Perhaps research using task inventories, which focus more on specific duties carried out by an incumbent, would indicate more clearly that the source of job analysis ratings is important, and that naive raters can never provide data comparable to that of expert raters when task inventories are used.

The potential comparability of naive and expert raters might also be further reduced by more extensive training of job analysts. Although care was taken in the present study to ensure that all of the raters were familiar with job analysis procedures in general, and the PAQ specifically, this "training" was surely not comparable to that received by job analysts in the field. More comprehensive training, as well as more experience with the PAQ, may well produce wider divergence among ratings from different sources.

The lack of more comprehensive training and experience is therefore a reason for exercising some caution in interpreting the results of the present study. The relatively small number of raters per job tends to reduce reliability estimates; this represents another reason for caution. In addition to dealing with these shortcomings, future studies must attempt to operationally define job stereotypes. Such stereotypes have not been addressed explicitly, but the implication of research comparing ratings from expert and naive job analysts is that PAQ data from

expert analysts may provide little more than stereotypes about the jobs they are describing. Perhaps the number of items rated DNA could suggest the extent to which stereotypes were influencing job analysis ratings. However, the fact that a certain PAQ item does not apply can just as well be valid and useful information about the job. Furthermore, just as there might be stereotypes about activities that do not apply, there are surely stereotypes about activities that do apply; focusing only on DNA items ignores these latter stereotypes. It is critical that we have some picture of what job stereotypes "look like." This information takes on even more significance if one speculates that sex- and age-linked stereotypes about jobs also exist and that these could influence job analysis data.

For the present, however, we can only conclude that there are certain jobs for which the PAQ may not be capable of providing much useful information. In these cases, when the PAQ is used, expert and naive raters can provide comparable job analysis information, raising the possibility that the PAQ is describing only common knowledge or stereotypes about the jobs (a possibility originally suggested in the work of Smith & Hakel, 1979). This further suggests that some contingency approach to job analysis be adopted. That is, certain job analysis instruments should be used only with some jobs, but not with others. Perhaps analysts' individual differences should be considered in such an approach for us to match analysts with instruments and instruments with jobs.

Although expert raters were found to provide information about some jobs that is beyond information available from naive raters, enough questions remain to conclude that neither researchers nor practitioners should take the basic validity of job analysis data for granted. It is worth noting at this point that researchers in other areas have recently taken heed of similar problems. For example, leadership researchers have found that leader behavior descriptions may largely be determined by leader's prototypes held by followers (Lord, Foti, & DeVader, 1984). Also, researchers in job design have found that ratings on the core dimensions of the Job Diagnostic Survey (Hackman & Oldham, 1975) are as much a function of such factors as worker affect, personality, and available social cues as of the actual motivating potential of the job being rated (e.g., Brousseau & Prince, 1981; James & Jones, 1980; Oldham, Hackman & Pearce, 1976; O'Reilly & Caldwell, 1979). It is not yet clear whether such factors influence PAQ-type job analysis data or how extensive such influence might be.

References

- Brousseau, K. R., & Prince, J. B. (1981). Job-person dynamics: An extension of longitudinal research. *Journal of Applied Psychology*, 60, 59-62.
- Cornelius, E. T. (1987). Practical findings from job analysis research. In S. Gael (Ed.), *Handbook of job analysis*. New York: Wiley.
- Cornelius, E. T., DeNisi, A. S., & Blencoe, A. G. (1984). Expert and naive raters using the PAQ: Does it matter? *Personnel Psychology*, 37, 453-464.
- Department of Labor. (1972). *Handbook for analyzing jobs*. Washington, DC: U.S. Government Printing Office.
- Department of Labor. (1977). *Dictionary of occupational titles*. Washington, DC: U.S. Government Printing Office.
- Fine, S. A., & Wiley, W. W. (1971). *An introduction to functional job analysis*. Washington, DC: Upjohn Institute for Employment Research.

- Gulliksen, H. (1968). Methods for determining equivalence of measures. *Psychological Bulletin*, 70, 534-544.
- Hackman, J. R., & Oldham, G. R. (1975). Development of the Job Diagnostic Survey. *Journal of Applied Psychology*, 60, 159-171.
- Harvey, R. J., & Hayes, T. L. (1986). Monte Carlo baselines for interrater reliability correlations using the Position Analysis Questionnaire. *Personnel Psychology*, 39, 345-357.
- James, L. R., & Jones, A. P. (1980). Perceived job characteristics and job satisfaction: An examination of reciprocal causation. *Personnel Psychology*, 33, 118-147.
- Lord, R. G., Foti, R. J., & DeVader, C. L. (1984). A test of leadership categorization theory: Internal structure, information processing, and leadership perceptions. *Organizational Behavior and Human Performance*, 34, 343-378.
- McCormick, E. J., DeNisi, A. S., & Shaw, J. B. (1979). Use of the Position Analysis Questionnaire for establishing the job component validity of tests. *Journal of Applied Psychology*, 64, 51-56.
- McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. (1972). A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ). *Journal of Applied Psychology*, 56, 347-368.
- Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.
- Oldham, G. R., Hackman, J. R., & Pearce, J. L. (1976). Conditions under which employees respond positively to enriched work. *Journal of Applied Psychology*, 61, 395-403.
- O'Reilly, C. P., & Caldwell, D. F. (1979). Informational influence as a determinant of perceived task characteristics and job satisfaction. *Journal of Applied Psychology*, 64, 157-165.
- Robinson, D. D., Wahlstrom, W. W., & Mecham, R. C. (1974). Comparison of job evaluation methods: A policy capturing approach using the Position Analysis Questionnaire. *Journal of Applied Psychology*, 59, 633-637.
- Smith, J. E., & Hakel, M. D. (1979). Convergence among data sources, response bias, and reliability and validity of a structured job analysis questionnaire. *Personnel Psychology*, 32, 677-692.
- Taylor, L. R. (1978). Empirically derived job families as a foundation for the study of validity generalization: Study 1. The construction of job families based on the component and overall dimensions of the PAQ. *Personnel Psychology*, 31, 537-561.

Received July 14, 1986

Revision received October 18, 1986

Accepted October 9, 1986 ■

Job-Related Stress, Social Support, and Burnout Among Classroom Teachers

Daniel W. Russell

Center for Health Services Research, University of Iowa

Elizabeth Altmaier and Dawn Van Velzen

College of Education, University of Iowa

In this study we examined the effects of job-related stressful events and social support on burnout among teachers. We conducted a mail survey of a random sample of public school teachers in Iowa. Consistent with findings in previous research, teacher characteristics such as age, sex, and grade level taught were predictive of burnout. We also found that the number of stressful events experienced and social support were predictive of teacher burnout. Some evidence of the stress-moderating role of social support was also found. Teachers who reported that they had supportive supervisors and indicated that they received positive feedback concerning their skills and abilities from others were less vulnerable to burnout. We discuss the implications of these findings for programs aimed at preventing teacher burnout.

Teaching has been identified as a particularly stressful occupation (Cacha, 1981; Farber & Miller, 1981; Landsman, 1978; Paine, 1981). Negative aspects of the job such as disciplinary problems, student apathy, overcrowded classrooms, involuntary transfers, excessive paperwork, inadequate salaries, demanding or unsupportive parents, and lack of administrative support are among the stressors that confront teachers. As a result of these stressful aspects of teaching, burnout among teachers occurs, expressed in physical (e.g., headaches, peptic ulcers), psychological (e.g., depression, anger), and behavioral (e.g., deterioration in work performance, absenteeism) symptoms (Cunningham, 1982). Teacher burnout is thought to be one reason for the increasing numbers of competent teachers who are leaving the classroom for alternative careers (Cunningham, 1982; Farber & Miller, 1981).

Empirical studies of teacher burnout have attempted to identify the teacher characteristics that are associated with higher burnout levels. Findings have indicated that age, sex, and grade level taught are significant predictors of scores on the Maslach Burnout Inventory (MBI; Anderson & Iwanicki, 1984; Beck & Gargiulo, 1983; Crane & Iwanicki, 1983; Schwab & Iwanicki, 1982b; Schwab, Jackson, & Schuler, 1984). Greater emotional exhaustion was reported by younger teachers. More negative attitudes toward students (depersonalization) were reported by male and secondary teachers. Finally, a greater sense of personal accomplishment was reported by elementary school teachers.

Although these relationships were statistically significant, note that these teacher characteristics accounted for less than 11% of the variance in burnout scores across these studies.

Social support has been identified as a resource that enables individuals to cope with stress (House, 1981). According to the moderating hypothesis, individuals who have supportive social relationships are able to rely on others to aid them in dealing with stressful situations. As a result, stress does not have negative effects on their physical and psychological health. On the other hand, individuals who lack supportive social relationships are vulnerable to the effects of stress. Empirical evidence regarding the moderating hypothesis is mixed. Some studies have indicated that social support interacts with stress in predicting physical and mental health, whereas other studies have not found the Stress \times Social Support interaction (for a recent review, see Cohen & Wills, 1985). However, researchers have consistently found that individuals who possess high levels of social support are in better physical and mental health.

Various writers (Kirk & Walter, 1981; Moracco & McFadden, 1982; Paine, 1981; Wangberg, 1982) have suggested that increasing the social supports available to teachers may be a useful strategy for preventing teacher burnout. However, empirical evidence concerning the impact of social support on burnout among teachers is very limited. A study of correctional teachers indicated that teachers who were classified as *burned out* spent less time with their fellow workers than did other teachers (Belcastro, Gold, & Grant, 1982). Zabel and Zabel (1982) reported that special education teachers who perceived greater administrative, peer, and parental support were less burned out. Finally, a survey of a random sample of school teachers in New Hampshire found that higher levels of social support from colleagues was associated with lower levels of burnout (Schwab et al., 1984).

The available evidence therefore suggests that programs designed to enhance the social supports available to teachers may represent an effective strategy for preventing teacher burnout. However, more detailed information concerning the impact of social support on teacher burnout is necessary before interven-

This research was supported by a bloc grant from the College of Education at the University of Iowa to Elizabeth Altmaier and Daniel Russell, and by Grant AG03846 from the National Institute on Aging to Daniel Russell and Carolyn Cutrona. A report of this research was presented at the American Psychological Association Convention in Toronto, Ontario, Canada, August 1984.

The authors would like to thank Carolyn Cutrona for her comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Daniel Russell, Center for Health Services Research, S 517 WL, University of Iowa, Iowa City, Iowa 52242.

tion programs can be developed. Consistent with studies of job stress and social support in other occupations (see reviews by House, 1981; Kasl & Wells, 1985), previous research on teacher burnout indicated that support received from co-workers and supervisors may be particularly important in preventing teacher burnout. In addition to understanding which individuals from the teacher's social network serve as important sources of support, we also need to determine what particular forms of social support are most useful in preventing teacher burnout (Cohen & Wills, 1985). For example, is having someone to turn to for advice in dealing with stressful situations effective in assisting teachers, or is deriving a general sense of competence from relationships with others the most important form of support?

The purpose of our research was to examine the impact of different facets of social support on teacher burnout. Assessments were made of social support received from different people in the teacher's social network (e.g., co-workers, spouses) and of the extent to which the teacher was receiving different forms of support from his or her relationships with others. Using these data, we examined the impact of both job-related stresses and social support on teacher burnout.

Method

Sample

Questionnaires were mailed to a stratified random sample of 600 public school teachers in Iowa during the final month of the academic school year. To enhance the representativeness of the sample, we sent 60% of the questionnaires to elementary school teachers and 40% to teachers who were teaching at the secondary level. We received back 316 completed questionnaires, a 53% return rate. Because it was the end of the school year, we could not conduct follow-up mailings to increase the return rate. However, this return rate is generally considered to be adequate for a mail survey (Babbie, 1973) and is comparable to the return rates reported in previous surveys of teacher burnout (Schwab & Iwanicki, 1982a, 1982b; Schwab et al., 1984).

The sample appeared to be generally representative of teachers in Iowa. Comparisons of the sample to all teachers in Iowa indicated that the sample was representative in terms of age ($M = 39.98$, $SD = 10.91$), years of teaching experience ($M = 14.75$, $SD = 8.84$), and proportion of the teachers who possessed graduate degrees (26.4%). A sampling bias did occur in terms of sex. Fewer male teachers returned our questionnaire (29.7%) than would be expected considering the proportion of male teachers in Iowa (39.9%), $\chi^2(1, N = 316) = 13.52$, $p < .001$.

Measures

Job-related stress. To develop a measure of job-related stress for teachers, we conducted a pilot study to identify the types of stressful events that teachers experience on the job. We asked a sample of 31 primary and secondary teachers to list the 3 most stressful events they had experienced at work during the previous school year. By combining similar events described by different teachers, a final list of 47 stressful events was constructed. These events appeared to be a representative sampling of the types of stressful situations teachers experience at work and ranged from problems with students and parents (e.g., "a student 'talks back' and argues"; "a parent does not make a student do homework") to interpersonal conflicts with other teachers and school administrators (e.g., "another teacher is inconsiderate while you are teaching"; "the school administration adds more areas/courses to teach").

Respondents to the mail survey were asked to indicate whether they had experienced each of the 47 stressful events during the current school year. If they had experienced an event, they rated how stressful the event was using a 0 to 7 scale, with 0 indicating that the event was not stressful at all and 7 indicating that the event was the most stressful the teacher had ever experienced. From these responses, two different stress measures were derived for each respondent: (a) the number of stressful events the teacher had experienced during the school year and (b) the sum of the subjective stress ratings by the teacher regarding those events he or she had experienced.

Social support. Two different social support measures were included in the mail questionnaire. First, a measure developed by House and Wells (1978) that focuses on support received from different members of the social network in the context of job-related stress was administered. This measure is designed to yield indexes of social support from four sources: supervisors, co-workers, spouses, and friends or relatives. Respondents are asked to rate how helpful each of these persons is in the context of work-related stresses. This measure appears to be quite reliable. Alpha coefficients ranging from .75 to .92 have been reported for the four subscales (House, 1981). Validity evidence for the scale has been provided by a number of studies conducted at the University of Michigan Institute for Social Research. Social support (particularly from supervisors and co-workers) has consistently been found to be predictive of measures of physical and mental health among workers in a variety of occupations (see House, 1981, for a review).

Respondents also completed the Social Provisions Scale (Russell & Cutrona, 1986). This measure is designed to assess the extent to which the person's current social relationships provide the six relational provisions described by Weiss (1974). These social provisions are (a) attachment, provided by intimate relationships in which the person receives a sense of security and safety; (b) social integration, provided by a network of social relationships in which individuals share interests and concerns; (c) reassurance of worth, provided by relationships in which the person's skills and abilities are acknowledged; (d) guidance, provided by relationships with trustworthy and authoritative individuals who can provide advice; (e) reliable alliance, derived from relationships in which the person can count on others for assistance under any circumstances; and (f) opportunity for nurturance, derived from relationships in which the person is responsible for the well-being of another.

Previous research has supported the reliability of the Social Provisions Scale. Alpha coefficients ranging from .66 to .76 for the subscales measuring each social provision have been found (Cutrona & Russell, in press). Validity for the measure has been provided by significant relations between social provision scores and measures of the individual's social network, satisfaction with different types of social relationships, and loneliness. Other research has indicated that scores on the Social Provisions Scale are predictive of adaptation to stressful situations (for a review, see Cutrona & Russell, in press). Finally, in an evaluation of the discriminant validity of the Social Provisions Scale, Russell and Cutrona (1986) found that scores on the measure were related to depression, independent of a variety of variables (e.g., social desirability, introversion-extraversion, neuroticism) that have been suggested in the literature as possible confounding factors.

Burnout. To assess teachers' reactions to job-related stress, the Maslach Burnout Inventory (Maslach & Jackson, 1981) was administered. The MBI consists of 22 items that are divided into three subscales (Emotional Exhaustion, Depersonalization, and Personal Accomplishment) that reflect different aspects of the burnout syndrome. Each item is rated on both an intensity and a frequency dimension. Scores on each subscale appear to be very reliable. Maslach and Jackson (1981) reported alpha coefficients ranging from .71 to .90 for the three subscales. Supporting the validity of the measure, burnout scores have been found to increase in stressful job settings and to predict job turnover and absenteeism (Maslach, 1982). For the purposes of this study, the reference

to *patients* in several of the items was changed to *students* to make the items relevant to teachers. Iwanicki and Schwab (1981), in a study of teachers, reported evidence supporting the reliability and validity of the MBI after the items had been modified to refer to students. Respondents in this study rated each item on the frequency dimension only, inasmuch as previous research has indicated that the intensity and frequency ratings are highly correlated (see Constable & Russell, 1986; Iwanicki & Schwab, 1981; Maslach & Jackson, 1981).

Results

Analyses of the data focused on several different issues. First, relations between teacher characteristics and the experience of job-related stress and burnout were evaluated. A final set of analyses examined the relation between job-related stress, social support, and burnout.

Teacher Characteristics, Stress, and Burnout

As described earlier, stress scores were computed for individual teachers based on both the number of events reported by each teacher and the sum of the stressfulness ratings made by the teacher for the events he or she had experienced. Teachers reported experiencing an average of 16.45 events during the previous academic school year ($SD = 8.08$). The number of events reported by the teacher was found to be highly correlated with the sum of stressfulness ratings, $r(314) = .83, p < .001$. Because of the high correlation between these measures and the similarity of the results using the two measures, the number of events reported by each teacher was used as the measure of job-related stress in subsequent analyses.

Regression analyses were conducted to examine the relation between teacher characteristics and reports of job-related stress and burnout. In conducting these analyses, a set of sociodemographic variables (i.e., sex, marital status, age, and community size) and variables related to the teacher's job (i.e., education, years of teaching experience, grade level taught, size of school, and average class size) were regressed on the number of stressful events reported by teachers and burnout scores from the MBI.

The results of these analyses are shown in Table 1. A very weak relation was found between teacher characteristics and reports of job-related stress, accounting for 6% of the variance in stress scores. The only predictor variable that was significantly related to the number of stressful events was age. A greater number of job-related stressors was reported by younger teachers.

Scores on the MBI were more strongly related to teacher characteristics. Overall, the predictor variables explained from 8.6% to 19.3% of the variance in burnout scores. As can be seen in Table 1, the results varied depending on the dimension of burnout. For emotional exhaustion, the statistically significant predictors were age and average class size. Greater emotional exhaustion was reported by younger teachers and by teachers who taught larger classes. Depersonalization was related to the teacher's sex and the grade level that he or she taught (elementary vs. secondary). Male teachers and teachers who taught the secondary grades reported higher levels of depersonalization. Marital status and grade level were significant predictors of scores on personal accomplishment. Teachers who were mar-

ried and who taught at the primary level reported greater feelings of personal accomplishment.

Stress, Social Support, and Burnout

The next set of analyses examined the relation between job-related stress, social support, and burnout among teachers. Two separate hierarchical regression analyses were conducted, one for each of the social support measures (i.e., support from network members and the social provisions) that were included in the survey. Predictor variables were entered into the regression equations in the following order. First, to control for the impact of teacher characteristics on burnout, the sociodemographic and job-related variables were entered into the regression equation. Second, the number of stressful events reported by the teacher was entered into the equation. Third, the social support variables were entered as a block of predictor variables into the regression equation. For the first regression analysis, the social support received from the teacher's supervisor, co-workers, spouse, and friends or relatives were used as predictors. For the second regression analysis, scores on the six social provisions from the Social Provisions Scale were used as predictors. Finally, in order to test for interactions between the stress and social support measures in predicting burnout, product terms were created by standardizing the stress and social support measures and multiplying them together. These product terms were then entered into the regression equations as a block of predictor variables.

After controlling for the effects of teacher characteristics on burnout, the number of stressful events reported by teachers was found to significantly predict emotional exhaustion, $R^2 = .088, F(1, 272) = 28.89, p < .001$, and depersonalization, $R^2 = .071, F(1, 272) = 26.22, p < .001$. As expected, the direction of these relations indicated that teachers who experienced greater stress also reported greater emotional exhaustion and depersonalization. No relation was found between job-related stress and personal accomplishment, $R^2 = .002$.

Regression results for the first set of social support measures, involving support received from supervisors, co-workers, spouses, and friends or relatives, are shown in Table 2. These social support measures explained from 5.0% to 6.3% of the variance in burnout scores, over and above the effects of teacher characteristics and job-related stress on burnout. Social support received from supervisors was found to be the only significant predictor of burnout. Teachers with supportive supervisors reported less emotional exhaustion, more positive attitudes toward students, and greater personal accomplishment.

Supervisor support was also found to interact with job-related stress in predicting depersonalization, $\beta = -.120, F(1, 264) = 4.90, p < .05$. The pattern of results associated with this interaction term were consistent with the buffering hypothesis. As the level of supervisor support increased, the strength of the relation between job-related stress and feelings of depersonalization decreased.

Table 3 reports the regression results for the six social provisions. Scores on the Social Provisions Scale explained from 8.6% to 14.6% of the variance in burnout scores after controlling for the effects of teacher characteristics and job-related stress on burnout. The most important social provision was re-

Table 1
Relations Between Teacher Characteristics and Job-Related Stress and Burnout

Teacher characteristic	Job-related stress		Emotional exhaustion		Depersonalization		Personal accomplishment	
	β	F	β	F	β	F	β	F
Sex	-.011	< 1	.031	< 1	-.229	12.25***	.105	2.30
Marital status	.041	< 1	-.043	< 1	-.023	< 1	.135	5.43*
Age	-.273	5.89*	-.250	5.07*	-.205	3.85	.022	< 1
Community size	-.053	< 1	.029	< 1	-.060	< 1	-.091	1.81
Education	.014	< 1	.004	< 1	-.073	1.35	.059	< 1
Teaching experience	.173	2.19	.055	< 1	.078	< 1	.073	< 1
Grade level	.123	2.87	.083	1.34	.186	7.64**	-.178	6.24*
School size	.033	< 1	.063	< 1	.108	2.96	-.070	1.12
Class size	.046	< 1	.163	7.00**	-.005	< 1	.083	1.86
R^2	.060		.086		.193		.101	
$F(8, 275)$	1.95*		2.86**		7.29***		3.43***	

Note. $F(1, 275)$ in column headings.
* $p < .05$. ** $p < .01$. *** $p < .001$.

assurance of worth. Teachers who indicated that other people respected their skills and abilities reported less emotional exhaustion, more positive attitudes toward students, and greater personal accomplishment. Feelings of depersonalization were also significantly related to reliable alliance. Teachers who indicated there were people they could count on in an emergency reported less depersonalization.

As was found for supervisor support, statistically significant interactions between job-related stress and reassurance of worth, $\beta = -.136$, $F(1, 260) = 5.31$, $p < .05$, and reliable alliance, $\beta = -.147$, $F(1, 260) = 4.58$, $p < .05$, were found in predicting depersonalization. These results were also consistent with the moderating hypothesis. As the level of reassurance of worth or reliable alliance increased, the strength of the relation between job-related stress and feelings of depersonalization decreased.

Discussion

Three aspects of the social support received by teachers, involving support from their supervisor, reassurance of worth,

and reliable alliance, were found to be predictive of burnout. Investigators in studies of other occupational groups such as nurses (Constable & Russell, 1986) and factory workers (House, 1981) have also found that supervisor support has positive effects on the physical and mental health of workers. These findings suggest that supervisory personnel should be the focus of programs designed to increase the social supports available to teachers. The relation between reassurance of worth and burnout provides guidance concerning the content of such a social support intervention. Acknowledgment of the teacher's skills and abilities by supervisory personnel would appear to be an important component to include in programs designed to prevent teacher burnout.

Reliable alliance also emerged as an important dimension of social support in relation to feelings of depersonalization. This form of support involves having people available in your social network to whom you can turn for assistance in an emergency. According to Weiss (1974), kin relationships are typically the source of this form of support. Cutrona and Russell (in press) have found that both friend and family relationships serve as a source of this provision. Thus, these results suggest that rela-

Table 2
Relation Between Social Support From Network Members and Burnout

Source of support	Emotional exhaustion		Depersonalization		Personal accomplishment	
	β	F	β	F	β	F
Supervisor	-.185	10.51**	-.126	5.37*	.152	6.40*
Co-worker	-.115	3.85	-.049	< 1	.105	2.91
Spouse	-.114	< 1	-.200	3.19	.116	< 1
Friend or relative	-.029	< 1	-.100	3.32	.070	1.34
R^2	.063		.050		.056	
$F(3, 268)$	5.56***		4.88***		4.49**	

Note. $F(1, 268)$ in column headings.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 3
Relation Between Social Provisions and Burnout

Provision	Emotional exhaustion		Depersonalization		Personal accomplishment	
	β	F	β	F	β	F
Attachment	-.032	< 1	-.116	2.35	-.043	< 1
Social integration	-.059	< 1	.144	3.85	.017	< 1
Reassurance of worth	-.206	10.81**	-.280	23.77***	.305	23.44***
Reliable alliance	-.145	3.52	-.170	5.76*	-.029	< 1
Guidance	.051	< 1	.052	< 1	.106	1.54
Opportunity for nurturance	.087	1.83	.030	< 1	.117	3.30
R^2	.086		.114		.146	
$F(5, 266)$	5.13***		8.08***		8.58***	

Note. $F(1, 266)$ in column headings.

* $p < .05$. ** $p < .01$. *** $p < .001$.

tionships outside the workplace may also play a role in assisting teachers in coping with job-related stress.

In a recent review of social support research, Cohen and Wills (1985) argued that evidence for a moderating effect of social support will occur when there is a correspondence between the type of support and the stressor that is affecting the individual. Our results concerning the stress-buffering effects of supervisor support and reassurance of worth on feelings of depersonalization would appear to be consistent with this theoretical perspective. Because the stressors being studied here all involved work-related situations, there would appear to be a clear correspondence between the nature of the stressor and the source (i.e., supervisor) and type of support (i.e., bolstering the teacher's self-esteem) that had stress-moderating effects.

Age was the only teacher characteristic that was found to be related to the amount of stress reported by teachers, with younger teachers indicating they had experienced a greater number of stressful events. This finding appears to reflect a generational bias in reporting stressful events. Older individuals may be reticent to admit to stressful situations, particularly those involving interpersonal conflicts. Other studies of stressful life events have also found that older individuals report lower levels of stress (Masuda & Holmes, 1978). An interesting topic for future research is an examination of how age is related to the experience of job-related stress among teachers, to evaluate whether this age difference reflects real differences in stressful experiences or a bias in the recall of events.

The age and sex of the teacher along with the grade level taught (primary vs. secondary) were found to be predictive of burnout scores. The pattern of these results, in terms of relations between teacher characteristics and the three dimensions of burnout, was identical to that reported by previous researchers (Anderson & Iwanicki, 1984; Beck & Gargiulo, 1983; Crane & Iwanicki, 1983; Schwab & Iwanicki, 1982b; Schwab et al., 1984). We also found marital status and average class size to be predictive of scores on personal accomplishment and emotional exhaustion, respectively. These findings indicate the types of teachers and teaching situations in which burnout is most likely to occur. One possible application of such findings lies in

targeting intervention programs to those most at risk for burnout.

It is important to emphasize that the data from this study were correlational, gathered at one point in time. We therefore cannot infer from our findings that job-related stress and social support are causally related to teacher burnout. An important area for future research concerns designing and carefully evaluating the effects of social support intervention programs in preventing teacher burnout. Such studies would provide information concerning possible causal relations between job-related stress, social support, and burnout. The findings from this study suggest the form that such intervention programs should take, indicating the most appropriate focus for such an intervention (i.e., the supervisor) and the forms of support that should be provided to teachers (i.e., reassurance of worth).

References

- Anderson, M. B., & Iwanicki, E. F. (1984). Teacher motivation and its relationship to burnout. *Educational Administration Quarterly*, 20, 109-132.
- Babbie, E. R. (1973). *Survey research methods*. Belmont, CA: Wadsworth.
- Beck, C. L., & Gargiulo, R. M. (1983). Burnout in teachers of retarded and nonretarded children. *Journal of Educational Research*, 76, 169-173.
- Belcastro, P. A., Gold, R. S., & Grant, J. (1982). Stress and burnout: Physiologic effects on correctional teachers. *Criminal Justice and Behavior*, 9, 387-395.
- Cacha, F. B. (1981). Teacher burnout: Causes and solutions. *Kappa Delta Pi Record*, 18, 23, 26-27.
- Cohen, S., & Wills, T. A. (1985). Stress, social support, and the buffering hypothesis. *Psychological Bulletin*, 98, 310-357.
- Constable, J. F., & Russell, D. (1986). The effect of social support and the work environment upon burnout among nurses. *Journal of Human Stress*, 12, 20-26.
- Crane, S., & Iwanicki, E. F. (1983, April). *The effect of role conflict and role ambiguity on perceived levels of burnout among special education teachers*. Paper presented at the meeting of the American Educational Research Association, Montreal.
- Cunningham, W. G. (1982). Teacher burnout: Stylish fad or profound problem. *Planning and Changing*, 12, 219-244.

- Cutrona, C. E., & Russell, D. (in press). The provisions of social relationships and adaptation to stress. In W. H. Jones & D. Perlman (Eds.), *Advances in personal relationships* (Vol. 1). Greenwich, CT: JAI Press.
- Farber, B. A., & Miller, J. (1981). Teacher burnout: A psychoeducational perspective. *Teachers College Record*, 83, 235-243.
- House, J. S. (1981). *Work stress and social support*. Reading, MA: Addison-Wesley.
- House, J. S., & Wells, J. A. (1978). Occupational stress, social support, and health. In A. McLean, G. Black, & M. Colligan (Eds.), *Reducing occupational stress: Proceedings of a conference* (Publication 78-140, pp. 8-29). Washington, DC: National Institute of Occupational Safety and Health.
- Iwanicki, E. F., & Schwab, R. L. (1981). A cross-validation study of the Maslach Burnout Inventory. *Educational and Psychological Measurement*, 41, 1167-1174.
- Kasl, S. V., & Wells, J. A. (1985). Social support and health in the middle years: Work and the family. In S. Cohen & S. L. Syme (Eds.), *Social support and health* (pp. 175-198). New York: Academic Press.
- Kirk, W., & Walter, G. (1981). Teacher support groups serve to minimize burnout: Principles for organizing. *Education*, 102, 147-150.
- Landsman, L. (1978). Is teaching hazardous to your health? *Today's Education*, 67(2), 48-50.
- Maslach, C. (1982). *Burnout—The cost of caring*. Englewood Cliffs, NJ: Prentice-Hall.
- Maslach, C., & Jackson, S. (1981). *Maslach Burnout Inventory manual*. Palo Alto, CA: Consulting Psychologists Press.
- Masuda, M., & Holmes, T. H. (1978). Life events: Perceptions and frequencies. *Psychosomatic Medicine*, 40, 236-261.
- Moracco, J. C., & McFadden, H. (1982). The counselor's role in reducing teacher stress. *The Personnel and Guidance Journal*, 60, 549-552.
- Paine, W. S. (1981). The burnout phenomenon. *Vocational Education*, 56(8), 30-33.
- Russell, D., & Cutrona, C. E. (1986). *The Social Provisions Scale: A multidimensional measure of perceived social support*. Manuscript in preparation.
- Schwab, R. L., & Iwanicki, E. F. (1982a). Perceived role conflict, role ambiguity, and teacher burnout. *Educational Administration Quarterly*, 18, 60-74.
- Schwab, R. L., & Iwanicki, E. F. (1982b). Who are our burned out teachers? *Educational Research Quarterly*, 7(2), 5-16.
- Schwab, R. L., Jackson, S. E. & Schuler, R. S. (1984, April). *Educator burnout: Sources and consequences*. Paper presented at meeting of the American Educational Research Association, New Orleans.
- Wangberg, E. G. (1982). Helping teachers cope with stress. *Educational Leadership*, 39, 452-454.
- Weiss, R. (1974). The provisions of social relationships. In Z. Rubin (Ed.), *Doing unto others* (pp. 17-26). Englewood Cliffs, NJ: Prentice-Hall.
- Zabel, R. H., & Zabel, M. K. (1982). Factors in burnout among teachers of exceptional children. *Exceptional Children*, 49, 261-263.

Received September 15, 1986

Revision received November 10, 1986

Accepted December 1, 1986 ■

Relative Weight, Smoking, and Mental Health as Predictors of Sickness and Absence From Work

Katharine R. Parkes

Department of Experimental Psychology, University of Oxford, Oxford, England

This article reports a longitudinal study of relative weight, smoking, and mental health as predictors of medically certified sickness and unauthorized absence from work among student nurses ($N = 185$). Information about smoking, relative weight, and self-reports of somatic complaints and social dysfunction was obtained prior to the 33-month period over which sickness and absence were recorded. Multiple regression was used to test a predictive model relating absence to linear and quadratic components of relative weight, smoking, and symptom measures. A significant curvilinear relation between relative weight and absence was found, the form of which closely resembled the relation between relative weight and mortality; smoking showed an additive effect. A linear interaction between social dysfunction and relative weight was also found; particularly high levels of absence occurred among those of high relative weight who also reported high levels of social dysfunction. Analysis of sickness episodes confirmed the adverse effects of overweight and, to a lesser extent, of underweight and smoking. The findings are discussed in terms of medical, psychological, and psychosocial influences on sickness and absenteeism.

Numerous studies in the psychological literature have examined individual and organizational predictors of sickness and absence from work. Extensive reviews by Muchinsky (1977) and Steers and Rhodes (1978) cover much of the earlier literature in this area, and more recent studies include those by Breaugh (1981), Cheloha and Farr (1980), Clegg (1983), Hammer and Landau (1981), Jenkins (1985), and Watson (1981). In contrast, in the occupational health literature, studies of sickness absence have been concerned with environmental health hazards and, in particular, with smoking as a predictor of absence (e.g., Athanasou, 1975; Bass, 1980; Holcomb & Meigs, 1972; Janzon, Lindell, & Trelle, 1981; U. S. Public Health Service, 1979). However, body build (assessed by weight in relation to height) has attracted little attention in either the psychological or the medical literature relating to absence from work, in spite of the evidence that excess weight is associated with a wide range of health problems and also has psychological implications. The more general occupational consequences of relative weight have also been largely disregarded, although a recent study by Hendrix, Ovalle, and Troxler (1985) demonstrated its significance as a predictor of physiological responses to work stress. The present study examined the role of relative weight

in relation to sickness and absence from work, while also taking into account two other significant factors, smoking and mental health. The background to the study is outlined in the following sections.

Relative Weight, Disease, and Mortality

Two recent reports (National Institutes of Health, 1985; Royal College of Physicians, 1983) reviewed evidence linking obesity with a number of serious medical conditions. In particular, large-scale longitudinal studies have shown that excess weight is a significant predictor of cardiovascular disease and that this effect is independent of other known risk factors (Hubert, Feinleib, McNamara, & Castelli, 1983; Kannel & Gordon, 1974; Lew & Garfinkel, 1979; Waaler, 1984). Overweight has also been linked to other major illnesses, for instance, diabetes and some types of cancer, and to hypertension (Bonham & Brock, 1985; Larsson, Bjorntorp, & Tibblin, 1981; Lew & Garfinkel, 1979; Rimm, Werner, Van Yserloo, & Bernstein, 1975; Stamler, Stamler, Riedlinger, Algera, & Roberts, 1978). Consistent with these findings is the report that excess weight is associated with higher mortality rates (Royal College of Physicians, 1983).

However, the relation between relative weight and mortality is not linear, but takes a curvilinear form that can be represented as a quadratic function (Dyer, Stamler, Berkson, & Lindberg, 1975). Thus, both overweight and underweight are associated with an increase in mortality above the minimal level, although the risk is greatest for overweight individuals. The nature and interpretation of this curvilinear relation has been a source of some controversy; the optimal value of relative weight is usually found to be approximately 10% below the average value (Bray, 1985), but there are considerable discrepancies between studies (Dyer et al., 1975; Sorlie, Gordon, & Kannel,

This research was funded by the Health and Safety Executive, London. The author gratefully acknowledges their financial support.

Thanks are also due to John Matthews, Department of Biomathematics, University of Oxford, for statistical advice; to Davina Rendall for research assistance; and to the staff and students of the Barnet Area School of Nursing, and the Occupational Health Departments of the two hospitals concerned.

Correspondence concerning this article should be addressed to Katharine R. Parkes, Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford, OX1 3UD, England.

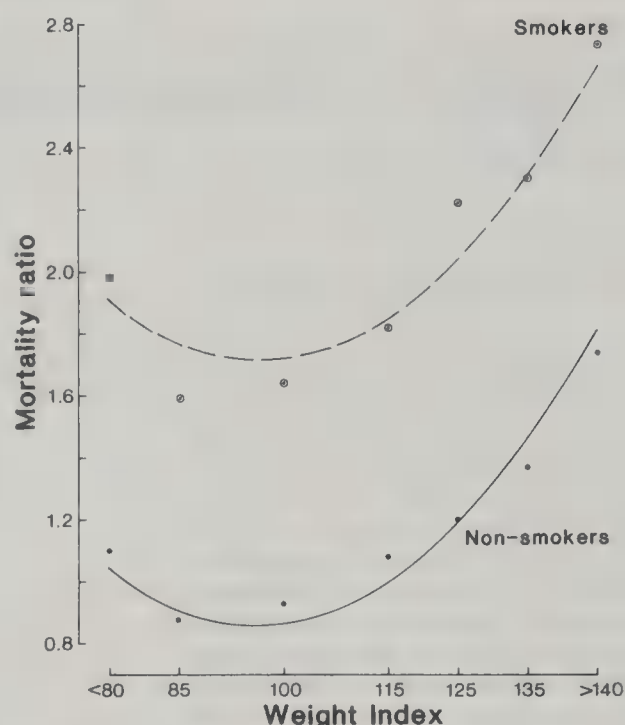


Figure 1. Mortality in relation to weight index for female smokers and nonsmokers. (The lines shown are quadratic curves fitted to data points reported by Lew and Garfinkel (1979). Each weight index value is the middle point of a relative weight range; 100 represents the sample mean.)

1980). Furthermore, some authors have attributed the upturn in the mortality curve at low weights to the confounding effect of smoking (Garrison, Feinleib, Castelli, & McNamara, 1983; Simopoulis, 1985). The possibility of confounding arises because smokers tend to have lower relative weight than do nonsmokers (Hawthorne & Murdoch, 1979; Khosla & Lowe, 1971; Kittel, Rustin, Dramaix, de Backer, & Kornitzer, 1978), but smoking is itself a major risk factor (U.S. Public Health Service, 1979). Therefore, failure to take smoking into account may distort observed relations between relative weight and mortality (Garrison et al., 1983). However, several studies have shown that the curvilinear relation between relative weight and mortality holds for both smokers and nonsmokers, the effect of smoking being independent of relative weight (Dyer et al., 1975; Lew & Garfinkel, 1979; Sorlie et al., 1980). Illustrative data for women are shown in Figure 1.

The majority of sickness episodes causing absence from work are due not to life-threatening illness, but to less serious health problems. Therefore, in the present context it is also necessary to consider the more general health implications of deviations from normal weight. Evidence suggests that excess weight (particularly severe obesity) may affect the normal functioning of major organ systems, including cardiovascular, respiratory, and endocrine systems (Royal College of Physicians, 1983). Furthermore, moderate or severe obesity causes damage to weight-bearing joints, and is associated with reduced physical fitness and work capacity (James, 1976). It would be predicted, therefore, that overweight would increase the risk not only of life-threatening diseases but also of less serious illnesses. Similarly, the underlying factors that link low weight to increased mortal-

ity may also be reflected in an increased vulnerability to more general health problems; in particular, evidence suggests that subclinical disease is a relevant factor (Sorlie et al., 1980), but a poor diet or previous illness may also be implicated.

Therefore, it is likely that not only major diseases but also lesser health problems will be more prevalent in overweight and underweight individuals than in those of average weight. If so, it would be predicted that the relation between relative weight and illness would be evident not only in episodes of prolonged absence from work but also in shorter episodes. Specifically, the present study tested the hypothesis that deviations from optimal weight would be manifest in increased frequency of sickness and absence from work, analogous to the increased risk of mortality and major disease. The existing literature provides little direct evidence of such effects; only two studies have examined relations between relative weight and absence from work, and neither have considered possible curvilinear effects. Zarling, Hartz, Larsen, and Rimm (1977) found that overweight was positively related to sickness absence, but not to unexplained absence. However, a longitudinal study by Larsson et al. (1981) found no significant effects of obesity on sickness absence.

Smoking Habits

In studying relative weight as a predictor of sickness absence, it is important to take into account the effects of smoking. Smokers show higher levels of absence from work than do nonsmokers, in terms of both number of episodes and time lost (Athanasou, 1975; Bass, 1980; Holcomb & Meigs, 1972; Janzon et al., 1981; U.S. Public Health Service, 1979). Also, as noted previously, smokers tend to have lower body weight than do nonsmokers. Thus, the problem of confounding is similar to that in the analysis of mortality data, and it is necessary to control for smoking in examining relationships between relative weight and absence. Indeed, failure to differentiate between smokers and nonsmokers may account for the nonsignificant findings of Larsson et al. (1981). In the present study, therefore, smoking behavior was included as a predictor variable, coded dichotomously. Additive effects of smoking and relative weight were predicted, consistent with findings relating to mortality data.

Mental Health

Mental health also influences sickness absence; several studies have demonstrated a relation between absence and psychoneurotic problems (Howell & Crown, 1971; Jenkins, 1980; Taylor, 1968; Tinning & Spry, 1981). More recently, Jenkins (1985) reported that, both retrospectively and prospectively, individuals classified as "cases" in a psychiatric interview had higher rates of absence—particularly, certificated sickness episodes—than did "normal" individuals. Similarly, high scores on Eysenck's neuroticism scale, a trait measure reflecting psychological vulnerability, are predictive of illness behavior, including bed disability days and physician consultations (Mechanic, 1980). It is possible that mental health and relative weight overlap as predictors of sickness absence, but existing evidence relating mental health to relative weight is ambiguous. For in-

stance, Bjorvell, Edman, Rossner, and Schalling (1985) found that obese individuals had high levels of somatic anxiety and muscular tension, whereas Crisp and McGuiness (1976) reported a negative relation between anxiety and obesity. Some studies have also shown obesity to be associated with high levels of extraversion (Hallstrom & Noppa, 1981) and low levels of neuroticism (Kittel et al., 1978), but an earlier review concluded that obesity was generally unrelated to neuroticism (Leon & Roth, 1977). Although the evidence is conflicting, the possibility that effects on sickness absence attributed to mental health may be confounded by variations in relative weight, and vice versa, cannot be ruled out. Therefore, in studying relative weight as a predictor of absence, it is desirable to take mental health into account.

Psychological Implications of Overweight

The implications of overweight and obesity are not only medical but also psychological and social. A slim physique is generally seen as desirable in Western culture, particularly for women (Leon & Finn, 1984); conversely, there is evidence that overweight individuals tend to be perceived negatively by others (Alton, 1975; DeJong, 1980; Herman, Zanna, & Higgins, 1986; Young & Powell, 1985). It is likely, however, that the psychological consequences of overweight have differential impact on those with different levels of mental health. Individuals who are overweight, but relatively free of neurotic symptoms, are likely to be less affected in either behavior or mood by the unfavorable social impact of their obesity. However, those who experience chronically high levels of psychological symptoms and are also overweight are more likely to be sensitive to the negative evaluation of others. Under such circumstances, withdrawal from the work situation may act as a means of coping with the combined effects of psychological distress, and the social disapproval and rejection encountered. Thus, interactive effects of relative weight and mental health, rather than solely additive effects, may predict absence. If so, those who combine overweight with poor mental health will show disproportionately frequent absences. However, underweight does not attract social disapproval in the way that overweight does; indeed, it may be socially endorsed. Therefore, interactions between mental health and relative weight would be expected to take a linear, rather than curvilinear form.

The Present Study

Predictions derived from the literature findings outlined in previous sections were tested using longitudinal data obtained from student nurses, for whom information about height, weight, smoking, and symptom levels was obtained at the time of enrollment in a 3-year hospital-based training course. To distinguish between somatic problems and affective distress, two self-report scales derived from the General Health Questionnaire (Goldberg & Hillier, 1979) assessing, respectively, somatic symptoms and social dysfunction were used. Episodes of absence (3 days or less in duration and not requiring a medical certificate) and sickness (more than 3 days, requiring medical certification) occurring over a 33-month period of training were

analyzed in relation to these predictor variables, using multivariate techniques. The predictive model included linear and curvilinear components of relative weight, smoking, somatic symptoms, and social dysfunction, and interactions between symptom measures and relative weight.

Method

Subjects

The subjects in this study were four annual intake groups (1977–1980) of female student nurses at a school of nursing in the London area. Participation was voluntary; 221 students agreed to take part, and 5 declined. Almost all of the students were in the 18–25-year-old age range, and were of British or Irish nationality.

Hospital Setting

For training purposes, the students were divided between two similar general hospitals. Prior to their first ward assignment, the students received 8 weeks of classroom instruction; they also had brief periods of classroom study during the 3 years of hospital-based training. Further details are given by Parkes (1982).

Data Collection

Height and weight. Each student was medically examined on arrival or shortly prior to arrival at the school of nursing. The medical record made at this time included height and weight (measured with light clothing, no shoes). These data were used to determine the relative weight for each subject, as described later.

Smoking. Information about smoking was obtained by self-report either prior to the initial period of nursing practice or during the 6th week of the initial ward assignment. The data were coded as 0 (non-smoker) or 1 (smoker). For students at one hospital, it was possible to check these self-reports against the medical records made on arrival; the close correspondence between the two sets of data indicated that self-report provided an accurate record of smoking behavior.

Measures of somatic symptoms and social dysfunction. The General Health Questionnaire, a symptom checklist developed by Goldberg (1978), was administered during the week immediately prior to the initial ward assignment. Subjects were asked to report symptoms experienced over the past month, and responses were scored on a 4-point Likert scale (0–3). Four 7-item subscales, which have been validated against clinical ratings (Goldberg & Hillier, 1979), can be derived from this questionnaire. In the present study, two of these subscales assessing somatic symptoms and social dysfunction, respectively, were used. The somatic symptoms scale assessed general lack of physical well-being and specific complaints such as headaches; the social dysfunction scale assessed psychological symptoms reflecting low morale, lack of confidence, and general dissatisfaction with life. These measures were chosen because they represented aspects of physical and psychological well-being that would be expected to show differential relations with sickness and absence.

Sickness and absence. Data relating to sickness and absence over the 3-year training period were obtained from the records of the school of nursing. This information was recorded as part of each student's formal training record; thus, a high level of accuracy was maintained. The data were classified into two categories. *Absence* referred to episodes of 1–3 days' duration for which medical certificates were not required; minor illnesses, particularly respiratory disorders, were the most frequent cause of these absences. If a student on the ward was obviously unwell, the senior nurse might informally excuse her from work. However, there

Table 1
Means, Standard Deviations, and Pearson Intercorrelations of Measures

Measure	M	SD	1	2	3	4	5	6
1. Relative weight	100.00	11.76	—					
2. Smoker–Nonsmoker ^a	0.37	0.48	−.03	—				
3. Somatic symptoms	4.77	3.32	.06	.12	—			
4. Social dysfunction	5.81	2.29	−.02	.10	.26**	—		
5. Absence ^b	3.48	1.49	.22**	.25**	.12	.08	—	
6. Height (in m)	1.63	0.06	.00	.05	.09	.08	.09	—
7. Weight (in kg)	57.50	7.91	.85**	−.01	.09	.03	.26**	.49*

Note. N = 185.
^a Smokers and nonsmokers were coded 1 and 0, respectively.
^b The absence scores are shown in their square-root transformed form.
* $p < .05$. ** $p < .01$.

was no formal check on the reasons for unexplained absences, particularly those of a single 1-day duration. *Sickness* referred to episodes that lasted more than 3 days, for which medical certificates had to be obtained.

For the purposes of the present study, frequency measures of absence were used. These measures are more reliable than time-lost measures and are recommended on psychometric grounds (Chadwick-Jones, Brown, Nicholson, & Sheppard, 1971; Hammer & Landau, 1981; Muchinsky, 1977). The number of absence and sickness episodes over a 33-month period, commencing 3 months after enrollment, was determined for each student. The 3-month delay ensured that the initial period of adaptation to the hospital environment was not included in the analysis and that the predictor variables were genuinely prospective. Students who left during the first year and those for whom data on any of the relevant variables were missing, were excluded. For students who completed at least 1 year but not the full 3-year training period, scores were determined proportionately from the frequency of episodes in the months completed. Data from 185 students were analyzed.

Calculation of Relative Weight

Two methods are available for calculating relative weight from height and weight data. One method depends on tables derived from mortality data. Using these tables, relative weight is expressed as the ratio of actual weight to ideal weight (i.e., the weight for which mortality is lowest) for a given height (see, e.g., Bray, 1985). Alternatively, relative weight can be calculated as a weight-for-height index without reference to mortality data; this method is usually preferred for research purposes and was used in the present study. The general form of such an index is W/H^p , where W represents weight, H represents height, and p is a power term (Benn, 1971). The criteria in determining p for a particular sample is that the index should show maximum correlation with weight and minimum correlation with height. The body mass index, which takes the form W/H^2 (expressed as Kg/m^2), is the most widely used index of this kind; it has been found to correlate highly with body fat measures (Garrow & Webster, 1985; Keys, Fidanza, Karvonen, Kimura, & Taylor, 1972). Following the recommendation of Benn (1971), the value of the exponential term p in the general weight/height index W/H^p was determined as the linear coefficient in the regression of Log W on Log H. The value of this coefficient was 2.04; thus, the body mass index (W/H^2) was an appropriate measure of weight-for-height in the present sample. The body mass index had a mean value of 21.69 ± 2.55 Kgs/ m^2 ; it correlated .85 with weight and .00 with height. For ease of interpretation and to obtain a scale-free measure, it was transformed to a measure of relative weight in the sample by dividing by the sample mean.

Statistical Treatment: Absence Scores

Data transformation. Absence data are typically positively skewed with a small proportion of very high values giving rise to a distribution in which the majority of scores lie below the mean. For scores of this kind, transforming the data will help to normalize the distribution. Examination of the absence data using the maximum likelihood methods of Box and Cox (1964) indicated that a square-root transformation was most appropriate in the present case. This transformation was therefore used prior to the regression analysis.

Regression analyses. The dependent measures in the regression analyses were the number of absence episodes occurring over the 33-month period following the initial 3 months of training, transformed as already described. The independent measures were entered hierarchically; main effects were entered prior to the product terms representing interactions (Cohen, 1978). The behavioral–objective measures, smoking and relative weight, were entered prior to the self-report measures. The order of entry was, therefore, smoking and relative weight (linear component), relative weight (quadratic component), somatic symptoms, and social dysfunction scores. Product terms representing interactions between somatic symptoms and relative weight and between social dysfunction and relative weight, were tested as the final term in the model. To facilitate the interpretation of main effects in the presence of interactions, the average-effect method advocated by Finney, Mitchell, Cronkite, and Moos (1984) was used; the relative weight values and the two symptom measures were transformed by subtracting their respective mean values prior to analysis. This transformation allows the first-order effects of two variables in the presence of an interaction between them to be interpreted as the effect of one variable at the mean value of the other.

Statistical Treatment: Sickness Scores

The distribution of certified sickness episodes was such that 30% of the group had no sickness during the 33-month period, a further 45% had 1 or 2 episodes, and the remaining 25% had between 3 and 20 episodes. It was not possible for a distribution of this type to use number of sickness episodes as the dependent variable in a regression analysis. An alternative approach was therefore adopted. The subjects were divided on the basis of sickness frequency into the three groups designated, and discriminant analysis was used to determine the extent to which smoking, relative weight, and symptom levels predicted group membership.

Results

Means, Standard Deviations, and Intercorrelations of Measures

The means, standard deviations, and intercorrelations of the independent measures (relative weight, height, smoking, so-

Table 2
Hierarchical Regression Analysis of Smoking, Relative Weight, Somatic Symptoms, and Social Dysfunction in Relation to Absence Episodes

Variable	Cumulative R^2	R^2 Increment	F	df	p	% of explained variance	B^a
Smoker-nonsmoker			13.35	1, 182	.0003		.830***
Relative weight (linear)	.114	.114	10.71	1, 182	.0013	69.51	.020*
+ Relative weight (quadratic)	.137	.023	4.84	1, 181	.029	14.02	.0014**
+ Somatic symptoms			0.62	1, 179	<i>ns</i>		.031
Social dysfunction	.144	.007	0.58	1, 179	<i>ns</i>	4.27	.050
+ Social Dysfunction \times Relative Weight	.164	.020	4.19	1, 178	.042	12.20	.0081*

Note. Plus (+) indicates a new step in the hierarchical analysis.

^a The B values are the unstandardized coefficients from the final regression equation in which each term is corrected for all other terms. The constant term in the final regression equation is 2.984.

* $p < .05$. ** $p < .025$. *** $p < .01$.

matic symptoms, and social dysfunction) and the dependent measure, absence, used in the regression analysis, are shown in Table 1. The mean number of absence episodes over the 33-month period was 14.33 ± 11.04 (mean duration = 1.44 days). For purposes of analysis, the raw data were square-root transformed (see Method section), giving a mean value of 3.48 ± 1.49 . The transformed absence scores are shown in Table 1 and are used in the following analyses.

As shown in Table 1, with the exception of somatic symptoms and social dysfunction ($r = 0.26$, $p < 0.001$), the independent measures were not significantly intercorrelated. In particular, smoking was unrelated to weight (mean relative weights of smokers and nonsmokers were 99.6% and 100.2%, respectively, of the overall group mean). However, there were several significant first-order correlations between absence and the independent variables; in particular, absence was positively related to relative weight and to smoking. Height and weight are also shown in Table 1 as individual measures, although only the combined relative weight measure was used in the predictive model. As would be expected, height and weight showed a highly significant positive correlation.

Regression Analyses: Absence Scores

Hierarchical analysis. The independent measures were entered in four hierarchical steps as predictors of absence: (a) smoking and relative weight (linear component), (b) relative weight (quadratic component), (c) somatic symptoms and social dysfunction, and (d) product terms representing the interactions of somatic symptoms and social dysfunction with relative weight. The interaction terms entered at Step (d) were tested separately as the final term in the model. At each stage, the variables entered were corrected for the effects of all variables entered above or at the same stage. The results of these analyses are shown in Table 2.

At the first stage of the analysis, the effects of smoking and relative weight were both significant when each was corrected for the other, and the model was highly significant overall, $F(2,$

182) = 13.71, $p < .0001$. At the second stage, the quadratic component of relative weight produced a significant increment to the multiple squared correlation, but the two symptom measures entered at the third stage were both nonsignificant. When entered as the final term in the model, the interaction between social dysfunction and the linear component of relative weight produced a significant increment in the squared multiple correlation, as shown in Table 2. However, the interaction between somatic symptoms and relative weight was not found to be significant when tested in a similar way, $F(1, 178) < 1$, *ns*.

At each stage of the analysis shown in Table 2, the model was highly significant overall. The multiple correlation at the final step was .405; the model accounted for 16.4% of the variance in absence scores. Examination of the successive squared multiple correlation values showed that smoking and relative weight (linear and quadratic components) jointly accounted for approximately 84% of the explained variance; the symptom measures and the interaction of somatic symptoms with relative weight accounted for the remaining 16%. The distribution of normalized residuals for the final model showed a close correspondence between the observed and the expected values throughout the range, indicating a good fit to the model.

Exploratory analyses. No interaction between smoking and relative weight had been predicted a priori, but the relevant interaction terms were tested on an exploratory basis to ensure that the additive effects shown in Table 2 did not obscure significant interaction terms. Both the linear interaction (Smoking \times Relative Weight) entered after the second step of the analysis and the curvilinear term (Smoking \times Relative Weight \times Relative Weight) entered after the linear term were found to be nonsignificant. In each case, the F value was less than 1.0.

A further exploratory analysis was carried out to determine the extent to which sickness acted as a predictor of absence. Frequency of sickness episodes was entered first into the regression equation (the skewed distribution did not preclude using sickness as an independent variable, but for consistency the

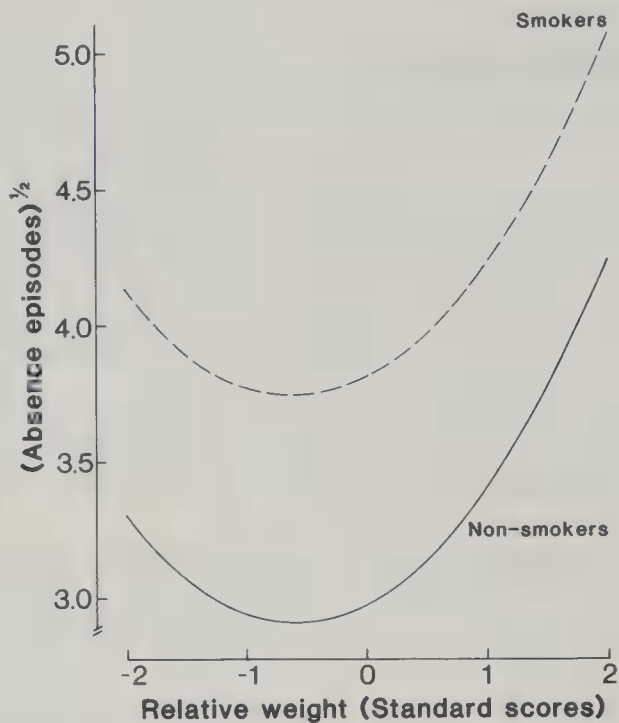


Figure 2. Effect of relative weight on absence for smokers and nonsmokers.

scores were square-root transformed). Sickness was found to be a highly significant predictor of absence, $F(1, 183) = 49.86$, $p < .0001$. The other variables were then entered in the same order as previously. The pattern of results was very similar to that shown in Table 2. In particular, relative weight and smoking remained significant predictors of absence after control for sickness: smoking, $p = .002$, and relative weight, $p = .055$ (linear) and $p = .034$ (quadratic). The overall proportion of variance accounted for by this model was 30.2%.

*Determination of Regression Coefficients:
Absence Scores*

To examine the nature and direction of main and interactive effects, regression coefficients were determined from a simultaneous regression analysis, including all the terms shown in Table 2. These coefficients are shown in the final column of Table 2. All of the regression coefficients were positive; smoking made a highly significant independent contribution to absence, whereas relative weight showed a significant curvilinear relation. Regression coefficients for the two symptom measures were nonsignificant, but the coefficient for the relative Weight \times Social Dysfunction interaction was significant. The unstandardized regression coefficients were used to derive the equation relating absence to relative weight and smoking habits. Terms involving symptom measures were not included in this equation as the relation between absence and relative weight was evaluated at mean symptom scores, which were transformed to zero prior to the regression analyses (see Method section). Relative weight was also transformed to a mean value of zero; its standard deviation was 11.758. The equation relating absence to relative weight follows.

$$\begin{aligned} & (\text{Absence episodes})^{1/2} \\ &= 0.020 \text{ RW} + 0.0014 \text{ RW}^2 + 0.830 \text{ S} + 2.984. \end{aligned}$$

In this equation, RW represents relative weight and S represents smoking habits (coded 1 for smokers and 0 for nonsmokers). This equation was used to derive the curves in Figure 2, in which the relation between absence and relative weight is shown separately for smokers and nonsmokers. The linear interaction between social dysfunction and absence was evaluated at the mean level of somatic symptoms, and regression lines were calculated for high relative weight (+1 SD above the mean) and for optimal relative weight (−0.6 SD below the mean). For illustrative purposes, the data for nonsmokers are plotted in Figure 3; for smokers the y-axis values are increased by .83.

Analysis of Sickness Episodes

As noted earlier in the article, the distribution of sickness episodes (medically certified episodes that lasted more than 3 days) was highly skewed, and consequently this measure could not be used as the dependent variable in a regression analysis. Instead, subjects were classified into three groups: no sickness episodes (Group 0, $n = 55$), 1–2 episodes (Group 1, $n = 84$), and 3–20 episodes (Group 2, $n = 46$). The proportions of smokers in the three groups were .25, .45, and .35 in Groups 0, 1, and 2, respectively. The difference between groups in the proportion of smokers fell just below the .05 level of significance, $\chi^2(2, N = 185) = 5.70$, $p = .058$. One-way analyses of variance showed that differences between the groups in mean relative weight were significant, $F(2, 182) = 3.95$, $p = .021$, that differ-

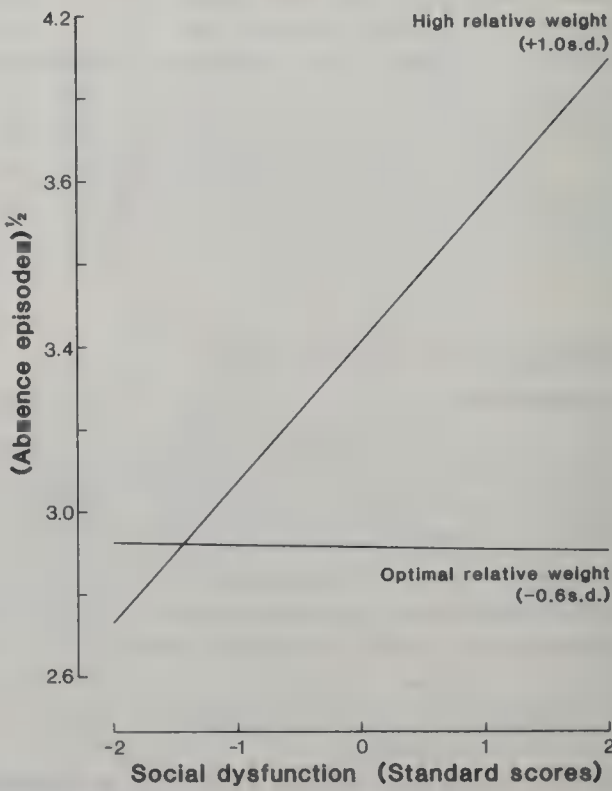


Figure 3. Effect of social dysfunction on absence for high and optimal levels of relative weight.

ences in somatic symptom scores were marginally significant, $F(2, 182) = 3.02, p = .051$, and that differences in social dysfunction scores were nonsignificant, $F(2, 182) = 0.46, ns$. Thus, the highest proportion of smokers was in Group 1, but Group 2 had the highest levels of relative weight and of somatic symptoms.

Discriminant analyses. The discriminant model used to predict group membership from the independent variables was analogous to that used in the regression analysis of absence scores except that no curvilinear term was included, as the dependent variable was nominal rather than continuous. The predictor variables were entered in two steps: first, smoking and relative weight, and second, somatic symptoms and social dysfunction. Pairs of variables entered simultaneously were corrected for each other. Rao's V (a generalized distance measure that evaluates the separation between the groups) was used to determine the significance of the model at each step. Use of Rao's V also enabled the significance of each variable entered to be determined, as the change in Rao's V for the entry of a single variable is distributed approximately as chi-square with $(g - 1)$ degrees of freedom, where g is the number of groups (see, e.g., Klecka, 1980). The magnitude of the change in Rao's V associated with the entry of a particular variable represents the relative importance of the variable in the model. It assesses the extent to which inclusion of the variable increases the separation between the groups.

The results of the analysis are shown in Table 3. At the first stage of the analysis, the predictive model was highly significant. Relative weight made a significant independent contribution ($p = .019$), but the effect of smoking just failed to reach the .05 level ($p = .054$). At the second step of the analysis, somatic symptoms and social dysfunction were entered into the discriminant model; only somatic symptoms made a significant contribution ($p = .046$). Neither of the product terms representing interactions between social dysfunction and relative weight and between somatic symptoms and relative weight—included as the final term in the model—produced a significant increment in Rao's V . Thus, the three significant predictors of group membership were smoking, relative weight, and somatic symptoms.

Evaluation of Discriminant Functions

To examine the effects of the predictor variables in more detail, the canonical discriminant functions were determined for the three-factor discriminant model. Two functions were derived: The first accounted for 69.9% of the total discriminating power of the two functions, and the second accounted for the remaining 30.1%.

The first discriminant function was highly significant, $\chi^2(6, N = 185) = 18.36, p = .005$. The canonical correlation, which is a measure of the association between the discriminant function and the dummy variables representing the three sickness groups, was .261. The square of this value represents the proportion of variance in the discriminant function that is explained by group membership—in this case 6.8%. For this function, the standardized discriminant coefficients for the predictor variables were .735 (relative weight), .635 (somatic

Table 3
Discriminant Analysis of Groups Differing in Sickness Frequency: Relative Weight, Smoking, Somatic Symptoms, and Social Dysfunction as Predictors of Group Membership

Variable	Rao's V	df	p	Rao's V		
				increment	df	p
Smoker-nonsmoker	13.72	4	.009	5.81	2	.054
Relative weight				7.93	2	.019
+ Somatic symptoms	21.44	8	.006	6.16	2	.046
Social dysfunction				2.43	2	.297

Note. $N = 185$. In the three groups defined by different frequencies of sickness, $n = 55$ (no sickness), $n = 84$ (1–2 episodes), and $n = 46$ (more than 2 episodes). Each term was corrected for terms entered at the same level or above. Plus (+) indicates the second step of the hierarchical analysis.

symptoms), and .142 (smoking). Thus, relative weight and somatic symptoms were the major determinants of the scores on this first discriminant function. When evaluated at the group means, the function value was highest for Group 2 (.425) and lowest for Group 0 (–.297); Group 1 was intermediate (–.038) but closer to Group 0 than Group 2. Thus, Group 2 was distinguished from Group 0 and Group 1 by high relative weight and high somatic symptoms, but these variables separated Group 0 and Group 1 only to a smaller extent. These results were further examined by repeating the discriminant analysis, with Groups 0 and 1 combined into a single group. In this analysis, only relative weight and somatic symptoms were significant predictors of group membership ($p = .0056$ and $p = .0387$, respectively); smoking was nonsignificant. Thus, membership of Group 2, as compared with Groups 0 and 1 combined, was associated with high relative weight and high somatic symptoms.

The second discriminant function in the main analysis fell just outside the .05 significance level, $\chi^2(2, N = 185) = 5.60, p = .060$. The canonical correlation value was .175; thus, for this function, 3.1% of the variance in scores was explained by group membership. Smoking was the major determinant of the function scores (discriminant coefficient, .953); relative weight made a negative contribution (–.264); and somatic symptoms contributed very little (.028). Group 1 had the highest value, .191; for both Group 0 and Group 2, the values were negative, –.187 and –.125, respectively. These results suggest that membership of Group 1, as compared with the other two groups, tended to be linked to smoking and to low relative weight but not to somatic symptoms. Overall, therefore, the results showed that high relative weight and high somatic symptoms distinguished those with high levels of sickness (Group 2) from those with no sickness (Group 0) or little sickness (Group 1). In addition, there was weaker evidence suggesting that Group 1 was distinguished from Group 0 and Group 2 by smoking and low relative weight.

Discussion

In demonstrating that relative weight is a highly significant predictor of sickness and absence from work, the present study

contributes new findings to a relatively unexplored area of research. Prospective data were used to test a model relating frequency of absence over a 33-month period to relative weight, while also taking into account smoking habits and self-reported symptom levels. The findings, both linear and curvilinear, were largely in line with a priori predictions. A more limited analysis of certificated sickness also demonstrated the influence of relative weight, symptoms, and to a lesser extent, smoking. These results are discussed in the following sections, with emphasis on the main and interactive effects of relative weight.

Characteristics of the Subject Group

The subjects of the present study were a relatively homogeneous population: All were female student nurses of comparable age and educational background; all were selected according to criteria that excluded applicants with a history of serious illness; and all were working in a similar environment. For each of the measures made, published data were available for comparison purposes.

Body mass. For the present subjects, mean body mass index (21.7 kgs/m^2) was slightly lower than the average value (22.7 Kgs/m^2) reported for a United Kingdom (U.K.) national sample of women aged 20–25 years (Royal College of Physicians, 1983) and was well within acceptable limits, $18.5\text{--}23.5 \text{ Kgs/m}^2$ (Bray, 1985). Of those outside the acceptable range, 20 students (11% of the group) were below the minimum value and 42 (23% of the group) were above the maximum.

Smoking. The percentage of smokers in the present subject group was 36.8%. This value corresponds closely to the proportion of smokers in the female population (Office of Population Censuses and Surveys, 1983) and to other reports of smoking among student nurses (Elkind, 1985; Hay, 1980; Hillier, 1981). However, in contrast to the findings reviewed at the beginning of this article, there was no difference in relative weight between smokers and nonsmokers in the present group, probably because smokers in the 18–25 age group would have had only a relatively short history of smoking.

General Health Questionnaire. As reported earlier (Parkes, 1982), the present group was comparable to the general female population in mean scores on the General Health Questionnaire. More specifically, the mean scores on the somatic and social dysfunction scales used in the present study were not markedly different from those reported by Banks (1983) for a young community sample.

Absence and sickness. The overall frequency of absence among subjects in the present study was relatively high, equivalent to 5.21 episodes per year. This rate of absence is approximately double that reported in hospital workers (Pines, Skulkeo, Pollak, Peritz, & Steif, 1985) and in other occupations (Breaugh, 1981; Clegg, 1983; Hammer & Landau, 1981). Furthermore, data for female U.K. employees (Jenkins, 1985) show a frequency of uncertificated absence episodes (of 1.3 days mean duration) less than one half that for the present group. Thus, the present data are consistent with other evidence that suggests that student nurses have high rates of short-term absence (Lunn, 1973; Menzies, 1960). However, the frequency of

certificated sickness in the present group (an average of less than one episode per year) was similar to Jenkins's (1985) data.

Relative Weight and Smoking as Predictors of Absence and Sickness Episodes

Absence. Consistent with a priori predictions, smoking and relative weight were significantly related to absence episodes. Jointly, these two variables explained 16.4% of the variance in absence; this proportion is more than three times that explained by work attitudes in a study by Breaugh (1981) and is comparable to that explained by a set of 10 demographic and work satisfaction variables in a study by Watson (1981). Thus, as compared with attitudinal and demographic variables, smoking and relative weight appear to be strong predictors of absence. For smoking, the results of the present study are consistent with existing evidence; however, the significance of relative weight as a predictor of absence and the curvilinear relation observed have not previously been reported. The relation between relative weight and absence frequency bears a strong resemblance to the mortality curves reproduced from the data of Lew and Garfinkel (1979) in Figure 1. This similarity extends to the position of the minimum point of the curve along the relative weight scale (93% of mean relative weight) and to the additive effect of smoking. The curvilinear relation is consistent with the view that all deviations from optimal weight are potentially detrimental (Simopoulis, 1985); the present findings show that these detrimental effects have direct implications for work attendance. The present data are also in accordance with mortality data showing that smoking is a more serious hazard than are all but the highest levels of overweight. Smokers of optimal weight had higher rates of absence than 90% of nonsmokers.

Sickness. The adverse effects of overweight were also apparent in the sickness data; high relative weight predicted membership of the high sickness group (Group 2). In addition, the second discriminant function suggested that low weight was associated with membership of Group 1 (1–2 sickness episodes); thus, both low and high relative weight tended to reduce the probability of membership of Group 0 (no sickness). The fact that the effects of relative weight on sickness were less clearly apparent than were the effects on absence can be attributed to the low frequency of sickness among the majority of the present subjects (who were young and had been selected on the basis of criteria that included good health) and the need to group the sickness data for analysis purposes.

Overweight and Underweight as Influences on Sickness and Absence

The present findings show that overweight and, to a lesser extent, underweight are potentially detrimental to well-being in ways that are manifest in absence from work. The concept of optimal body mass derives from data on mortality and major diseases, but the present study extends its relevance to less serious forms of disease and incapacity and to transient impairment of well-being. However, in spite of the similarities between the present findings and those from studies of serious illness, it should not be inferred that absence from work is solely due to

medical problems. As discussed earlier, body size also has important psychological and psychosocial implications. Thus, both medical and psychological effects may underly the findings.

The relation between overweight and sickness episodes suggests that one manifestation of the impairment in the functioning of major organ systems associated with obesity (James, 1976) is increased frequency of overt sickness or incapacity diagnosable by medical examination and necessitating medically sanctioned leave from work. Thus, excess weight is a risk factor for illnesses that are not life threatening but that interfere temporarily with an individual's capacity to fulfil normal work and social roles. However, the cause of short-term absence was less clear. One possibility is that absence might occur as an antecedent to, or a consequence of, specific episodes of more serious illness. However, this mediating effect did not explain the present results. Although sickness was correlated with absence, the inclusion of sickness in the regression model had very little effect on the pattern of results relating absence to relative weight and smoking. It is therefore necessary to consider, first, the possibility that some short-term health problems that impair the capacity to work may be linked to deviations from optimal weight without being manifest in more serious illness, and to consider, second, the role of psychological and psychosocial factors as influences on absence.

In regard to the first point, excess weight may render individuals more vulnerable to minor health problems such as respiratory disorders, which are the most common cause of lost time among nurses (Hillier, 1981; Lunn, 1973). It is also possible that overweight individuals experience a greater degree of physical incapacity as a result of minor respiratory disorders than do their normal-weight peers. Consistent with this is Shephard's (1977) suggestion that ability to stay at work during a minor illness is linked to cardiorespiratory fitness, which is impaired by obesity. Conversely, the increased frequency of absence among underweight individuals may be due to subclinical disease (Sorlie et al., 1980) or to nutritional deficiencies, both of which could reduce ability to meet the physical workload inherent in nursing and lower resistance to infection without necessarily giving rise to longer-term illness.

In regard to the second point, the persistence of the significant curvilinear relation between relative weight and absence after control for sickness could also be explained by psychological and psychosocial mechanisms. For instance, overweight individuals tend to be perceived negatively by others; thus, they may receive less social support at work than their normal-weight peers and, consequently, be more likely to respond to adverse circumstances by absence. The interaction between relative weight and social dysfunction, discussed later, lends support to this explanation. It is also possible that the link between absence and relative weight is indirectly mediated by individual differences in personality and by coping behaviors. These indirect effects could apply to both underweight and overweight individuals. For instance, certain personality traits such as impulsiveness and low need for social approval, may be associated with both deviations from normal weight and absenteeism. Similarly, maladaptive forms of coping, such as avoidance and denial, may be manifest in both weight problems and absence

from work; for instance, Staw (1977) has suggested that absenteeism represents an attempt to cope with work stress by withdrawal.

Symptom Levels as Predictors of Sickness and Absence

After taking into account relative weight and smoking, both somatic symptoms and social dysfunction were found to have significant implications for sickness and absence. Several aspects of the findings merit comment:

1. Initial level of somatic symptoms directly predicted sickness but was unrelated to absence. In contrast, social dysfunction predicted absence (but only among subjects of high relative weight) and was unrelated to sickness. These differential effects highlight the link between somatic complaints and sickness episodes as compared with the influence of low morale on absence behavior. Thus, the use of global measures of mental health to predict sickness and absence (e.g., Jenkins, 1980; Parkes, 1983; Tinning & Spry, 1981) may conceal more specific relations that link particular kinds of symptoms to different types of work incapacity.

2. Longitudinal data reported by Mechanic (1980) suggest that reports of physical complaints by adults are predicted by the experience of major illness and by illness behavior during childhood. Thus, to some extent, somatic symptom reports among the present subjects may reflect not only current somatic concerns but also past history. A common pattern of health problems and attitudes persisting from childhood through to adult life may therefore underly the link between high levels of somatic symptoms and subsequent sickness episodes observed in the present study.

3. The interaction between relative weight and social dysfunction predicted absence; only among subjects of high relative weight was low morale linked to withdrawal from the work situation in the form of absenteeism. These findings suggest that the capacity to maintain a body weight that is optimal from the point of view of physical health and work attendance may also be associated with a more general psychological resilience and psychosocial resources for managing stress. Physical exercise, which has been shown to moderate stress-strain relations (Kobasa, Maddi, & Puccetti, 1982; Roth & Holmes, 1985) may be one factor underlying this finding; exercise programs have been found to reduce weight and to reduce absenteeism (Cox, Shephard, & Corey, 1981). However, the social consequences of overweight are also relevant. Excess weight tends to be seen as deviant and to attract unfavorable responses from others (Allon, 1975); thus, individuals who are overweight may receive relatively little social support when they experience difficulties at work, and therefore they may be more prone to absence. Consistent with this explanation, coping by withdrawal is more common when the work environment is perceived as unsupportive (Newton & Keenan, 1985). There is also evidence that exercise and social support combine additively to decrease the likelihood of illness in response to stress (Ouellette Kobasa, Maddi, Puccetti, & Zola, 1985). More specifically, it is relevant that nursing is an occupation that demands high levels of physical activity and dexterity and imposes the additional strain of being in a very conspicuous role among patients in the wards. In all

of these respects, student nurses who were both overweight and low in morale at the start of training may have been at a particular disadvantage relative to normal-weight peers.

4. In a previous study, Parkes (1983) found that absence over a 6-month period was predicted by initial affective state only among those who reported themselves as likely to smoke under stress. Among nonsmokers and relaxation smokers, affective state did not predict absence. The form of the interaction was analogous to that between relative weight and social dysfunction in the present study; thus, both high relative weight and the tendency to use smoking as a means of coping with stress represent vulnerability factors that moderate the relation between psychological distress and absence. However, there was no evidence of confounding between the effects of type of smoking behavior and the effects of relative weight. In the present study, as noted earlier, no difference in relative weight was found between smokers and nonsmokers, and examination of previous data showed that type of smoking behavior (stress smoking vs. relaxation smoking) was also unrelated to relative weight.

General Points

The present study raises a number of methodological issues of general relevance to the study of sickness and absence. First, the study was longitudinal; information about the predictor variables was collected prior to the period during which sickness and absence was monitored. Thus, the possibility that the observed levels of relative weight, smoking, and symptoms were effects of rather than causes of the sickness and absence episodes occurring during the study period, can be ruled out. In contrast, inferences from retrospective studies in which the sickness and absence episodes analyzed occur prior to the research observations, are inevitably ambiguous as to causal direction. Furthermore, because the present data were collected during the subjects' initial occupational experience, previous job experience does not distort the findings. However, the possibility that antecedent factors of a constitutional or psychological nature may have influenced both dependent and independent measures cannot be excluded.

Second, the existing literature on absence from work has tended to focus on mental health, attitudinal, and demographic variables and has relied largely on bivariate methods of analysis. Steers and Rhodes (1978), proposing a process model of absence, emphasized the need for multivariate analysis techniques. Findings from the present study underline this point. Multivariate analyses, in which smoking and relative weight (both of which have direct implications for health and well-being) are controlled, could potentially enhance the sensitivity of analyses intended to test demographic, attitudinal, and affective predictors of absence. Third, smoking and relative weight tend to be more stable characteristics than do short-term psychological states. Thus, in a study covering several years, it would be expected that their continuing effects would be stronger than those of initial mental state. Consistent with this, somatic and psychological symptoms were less strongly correlated with absence scores over the 33-month period than were smoking and relative weight.

In view of the strong health implications of smoking and rela-

tive weight in men and women of all age groups, it would be expected that the findings of the present study would also be observed in other occupations; further longitudinal studies would serve to clarify this issue. If the present findings do apply more generally, then there is a strong case for encouraging organizational involvement in programs intended to facilitate maintenance of physical fitness and optimal weight among employees (see, e.g., Brownell, Stunkard, & McKeon, 1985; Cox et al., 1981).

References

- Allon, N. (1975). The stigma of overweight in everyday life. In G. A. Bray (Ed.), *Obesity in perspective* (Fogarty International Center Series of Preventive Medicine, Vol. 2, Pt. 2, pp. 83-102). Washington, DC: U.S. Government Printing Office.
- Athanasou, J. A. (1975). Sickness absence and smoking behavior and its consequences: A review. *Journal of Occupational Medicine*, 17, 441-445.
- Banks, M. H. (1983). Validation of the General Health Questionnaire in a young community sample. *Psychological Medicine*, 13, 349-353.
- Bass, F. (1980). Epidemiology of cigarettes: Usage, illness and cessation. *Behavioral Medicine Update*, 2, 13-16.
- Benn, R. T. (1971). Some mathematical properties of weight-for-height indices as measures of adiposity. *British Journal of Preventive Medicine*, 25, 42-50.
- Bjorvell, H., Edman, G., Rossner, S., & Schalling, D. (1985). Personality traits in a group of severely obese patients: A study of patients in two self-chosen weight reducing programs. *International Journal of Obesity*, 9, 257-266.
- Bonham, G. S., & Brock, D. B. (1985). The relationship of diabetes with race, sex and obesity. *American Journal of Clinical Nutrition*, 41, 776-783.
- Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society (B)*, 26, 211-252.
- Bray, G. A. (1985, April). Obesity: Definitions, diagnosis and disadvantages. *Medical Journal of Australia*, 142 (Special Suppl.), S2-8.
- Breaugh, J. A. (1981). Predicting absenteeism from prior absenteeism and work attitudes. *Journal of Applied Psychology*, 66, 555-560.
- Brownell, K. D., Stunkard, A. J., & McKeon, P. E. (1985). Weight reduction at the worksite: A promise partially fulfilled. *American Journal of Psychiatry*, 142, 47-52.
- Chadwick-Jones, J. K., Brown, C. A., Nicholson, N., & Sheppard, C. (1971). Absence measures: Their reliability and stability in an industrial setting. *Personnel Psychology*, 24, 463-470.
- Cheloha, R. S., & Farr, J. L. (1980). Absenteeism, job involvement, and job satisfaction in an organizational setting. *Journal of Applied Psychology*, 65, 467-473.
- Clegg, C. W. (1983). Psychology of employee lateness, absence, and turnover: A methodological critique and an empirical study. *Journal of Applied Psychology*, 68, 88-101.
- Cohen, J. (1978). Partialled products are interactions; partialled powers are curve components. *Psychological Bulletin*, 85, 858-866.
- Cox, M., Sheppard, R. J., & Corey, P. (1981). Influence of an employee fitness programme upon fitness, productivity, and absenteeism. *Ergonomics*, 24, 795-806.
- Crisp, A. H., & McGuinness, B. (1976). Jolly fat: Relation between obesity and psychoneurosis in general population. *British Medical Journal*, 1, 7-9.
- DeJong, W. (1980). The stigma of obesity: The consequences of naive assumptions concerning the causes of physical deviance. *Journal of Health and Social Behavior*, 21, 75-87.
- Dyer, A. R., Stamler, J., Berkson, D. M., & Lindberg, H. A. (1975).

- Relationship of relative weight and body mass index to 14-year mortality in the Chicago Peoples Gas Company study. *Journal of Chronic Disease*, 28, 109-123.
- Elkind, A. K. (1985). The social definition of women's smoking behaviour. *Social Science and Medicine*, 20, 1269-1278.
- Finney, J. W., Mitchell, R. E., Cronkite, R. C., & Moos, R. H. (1984). Methodological issues in estimating main and interactive effects: Examples from coping/social support and stress field. *Journal of Health and Social Behavior*, 25, 85-98.
- Garrison, R. J., Feinleib, M., Castelli, W. P., & McNamara, P. M. (1983). Cigarette smoking as a confounder of the relationship between relative weight and long-term mortality. The Framingham heart study. *Journal of the American Medical Association*, 249, 2199-2203.
- Garrow, J. S., & Webster, J. (1985). Quetelet's index (W/H^2) as a measure of fatness. *International Journal of Obesity*, 9, 147-153.
- Goldberg, D. (1978). *Manual of the General Health Questionnaire*. Windsor, England: NFER Publishing Company.
- Goldberg, D., & Hillier, V. F. (1979). A scaled version of the General Health Questionnaire. *Psychological Medicine*, 9, 139-145.
- Hallstrom, T., & Noppa, H. (1981). Obesity in women in relation to mental illness, social factors, and personality traits. *Journal of Psychosomatic Research*, 25, 75-82.
- Hammer, T. V., & Landau, J. (1981). Methodological issues in the use of absence data. *Journal of Applied Psychology*, 66, 574-581.
- Hawthorne, V. M., & Murdoch, R. M. (1979). Body weight of men and women aged 40-64 years from an urban area in the west of Scotland. *Community Medicine*, 1, 229-235.
- Hay, D. R. (1980). The smoking habits of nurses in New Zealand: Results from the 1976 population census. *New Zealand Medical Journal*, 92, 391-393.
- Herman, C. P., Zanna, M., & Higgins, E. T. (Eds.). (1986). Physical appearance, stigma, and social behavior. *Proceedings of the Third Ontario Symposium on Personality and Social Psychology*. Hillsdale, NJ: Erlbaum.
- Hendrix, W. H., Ovalle, N. K., & Troxler, R. G. (1985). Behavioral and physiological consequences of stress and its antecedent factors. *Journal of Applied Psychology*, 70, 188-201.
- Hillier, A. (1981, February). Stresses, strains and smoking. *Nursing Mirror*, pp. 26-30.
- Holcomb, H. S., & Meigs, J. W. (1972). Medical absenteeism among cigarette, cigar and pipe smokers. *Archives of Industrial Health*, 25, 295-300.
- Howell, R. W., & Crown, S. (1971). Sickness absence levels and personality inventory scores. *British Journal of Industrial Medicine*, 28, 126-130.
- Hubert, H. B., Feinleib, M., McNamara, P. M., & Castelli, W. P. (1983). Obesity as an independent risk factor for cardiovascular disease: A 26-year follow-up of participants in the Framingham heart study. *Circulation*, 67, 968-977.
- James, W. P. T. (1976). *Research on obesity: A report of the DHSS/MRC group*. London: Her Majesty's Stationery Office.
- Janzon, L., Lindell, S. E., & Trelle, E. (1981). Smoking and disease: Attitudes and knowledge in middle-aged men. *Scandinavian Journal of Social Medicine*, 9, 127-133.
- Jenkins, R. (1980). Minor psychiatric morbidity in employed men and women and its contribution to sickness absence. *Psychological Medicine*, 10, 751-757.
- Jenkins, R. (1985). Minor psychiatric morbidity and its contribution to sickness absence. *British Journal of Industrial Medicine*, 42, 147-154.
- Kannel, W. B., & Gordon, T. (1974). Obesity and cardiovascular disease: The Framingham study. In W. L. Burland, P. D. Samuel, & J. Yudkin (Eds.), *Obesity* (pp. 24-51). London: Churchill Livingstone.
- Keys, A. F., Fidanza, M. J., Karvonen, M. J., Kimura, N., & Taylor, H. L. (1972). Indices of relative weight and obesity. *Journal of Chronic Disease*, 25, 329-343.
- Khosla, T., & Lowe, C. R. (1971). Obesity and smoking habits. *British Medical Journal*, 4, 10-13.
- Kittel, F., Rustin, R. M., Dramaix, M., de Backer, G., & Kornitzer, M. (1978). Psycho-social biological correlates of moderate overweight in an industrial population. *Journal of Psychosomatic Research*, 22, 145-158.
- Klecka, W. R. (1980). *Discriminant analysis*. Beverly Hills, CA: Sage.
- Kobasa, S. C., Maddi, S. R., & Puccetti, M. C. (1982). Personality and exercise as buffers in the stress-illness relationship. *Journal of Behavioral Medicine*, 5, 391-404.
- Larsson, B., Bjorntorp, P., & Tibblin, G. (1981). The health consequences of moderate obesity. *International Journal of Obesity*, 5, 97-116.
- Leon, G. R., & Finn, S. (1984). Sex role stereotypes and the development of eating disorders. In C. S. Widom (Ed.), *Sex roles and psychopathology* (pp. 317-337). New York: Plenum Press.
- Leon, G. R., & Roth, L. (1977). Obesity: Psychological causes, correlations, and speculations. *Psychological Bulletin*, 84, 117-139.
- Lew, E. A., & Garfinkel, L. (1979). Variations in mortality by weight among 750,000 men and women. *Journal of Chronic Disease*, 32, 563-576.
- Lunn, J. A. (1973). Hospital hazards. *Practitioner*, 210, 490-499.
- Mechanic, D. (1980). The experience and reporting of common physical complaints. *Journal of Health and Social Behavior*, 21, 146-155.
- Menzies, I. E. P. (1960). A case-study in the functioning of social systems as a defence against anxiety. *Human Relations*, 13, 95-121.
- Muchinsky, P. M. (1977). Employee absenteeism: A review of the literature. *Journal of Vocational Behavior*, 10, 316-340.
- National Institutes of Health (1985). Health implications of obesity. *Annals of Internal Medicine*, 103, 147-151.
- Newton, T. J., & Keenan, A. (1985). Coping with work-related stress. *Human Relations*, 38, 107-126.
- Office of Population Censuses and Surveys (1983). *General household survey* (GHS 83/3). London: Office of Population Censuses and Surveys.
- Ouellette Kobasa, S. C., Maddi, S. R., Puccetti, M. C., & Zola, M. A. (1985). Effectiveness of hardiness, exercise and social support as resources against illness. *Journal of Psychosomatic Research*, 29, 525-533.
- Parkes, K. R. (1982). Occupational stress among student nurses: A natural experiment. *Journal of Applied Psychology*, 67, 784-796.
- Parkes, K. R. (1983). Smoking as a moderator of the relationship between affective state and absence from work. *Journal of Applied Psychology*, 68, 698-708.
- Pines, A., Skulkeo, K., Pollak, E., Peritz, E., & Steif, J. (1985). Rates of sickness absenteeism among employees of a modern hospital: The role of demographic and occupational factors. *British Journal of Industrial Medicine*, 42, 326-335.
- Rimm, A. A., Werner, L. H., Van Yserloo, V., & Bernstein, R. A. (1975). Relationship of obesity and disease in 73,532 weight-conscious women. *Public Health Report*, 90, 44-51.
- Roth, D. L., & Homes, D. S. (1985). Influence of physical fitness in determining the impact of stressful life events on physical and psychological health. *Psychosomatic Medicine*, 47, 164-173.
- Royal College of Physicians (1983). Obesity: A report of the Royal College of Physicians. *Journal of the Royal College of Physicians*, 17, 3-58.

- Shephard, R. J. (1977). *Endurance fitness* (2nd ed.). Toronto, Ontario, Canada: University of Toronto Press.
- Simopoulos, A. P. (1985). The health implications of overweight and obesity. *Nutrition Reviews*, 43, 33-40.
- Sorlie, P., Gordon, T., & Kannel, W. B. (1980). Body build and mortality. *Journal of the American Medical Association*, 243, 1828-1831.
- Stamler, R., Stamler, J., Riedlinger, W. F., Algera, G., & Roberts, R. H. (1978). Weight and blood pressure: Findings in hypertension screening of 1 million Americans. *Journal of the American Medical Association*, 240, 1607-1610.
- Staw, B. M. (1977). Motivation in organizations: Towards a synthesis and redirection. In B. M. Staw & G. R. Salancik (Eds.), *New directions in organizational behavior* (pp. 55-95). Chicago: St. Clair Press.
- Steers, R. M., & Rhodes, S. R. (1978). Major influences on employee attendance: A process model. *Journal of Applied Psychology*, 63, 391-407.
- Taylor, P. J. (1968). Personal factors associated with sickness absence: A study of 194 men with contrasting sickness absence experience in a refinery population. *British Journal of Industrial Medicine*, 25, 106-118.
- Tinning, R. J., & Spry, W. B. (1981). The extent and significance of stress symptoms in industry—with examples from the steel industry. In E. N. Corlett & J. Richardson (Eds.), *Stress, work design and productivity* (pp. 129-148). Chichester, England: Wiley.
- U. S. Public Health Service. (1979). *Smoking and health: A report of the Surgeon-General* (DHEW Pub. No. PHS 79-50066). Washington, DC: Author.
- Waler, H. Th. (1984). Height, weight and mortality: The Norwegian experience. *Acta Medica Scandinavica* (Suppl. 679), 1-56.
- Watson, C. J. (1981). An evaluation of some aspects of the Steers and Rhodes model of employee attendance. *Journal of Applied Psychology*, 66, 385-389.
- Young, L. M., & Powell, B. (1985). The effects of obesity on the clinical judgments of mental health professionals. *Journal of Health and Social Behavior*, 26, 233-246.
- Zarling, E., Hartz, A., Larsen, J., & Rimm, A. A. (1977). Obesity and illness associated absenteeism. *Obesity/Bariatric Medicine*, 6, 134-136.

Received May 19, 1986

Revision received September 29, 1986

Accepted December 1, 1986 ■

Effects of Role Loss on Work-Related Attitudes

Judith A. Schlenker and Barbara A. Gutek
Claremont Graduate School

In this field study, some professionals in a large social services agency were reassigned to nonprofessional jobs. These reassignments provided an opportunity to study the impact of work role-loss disassociated from the loss of employment and salary. The sample consisted of 132 government-employed social workers, one half of whom were reassigned to nonprofessional jobs. Data show that work role-loss is associated with lower job satisfaction ($p < .001$), lower work-related self-esteem ($p < .01$), and higher scores on the measure of intention to leave the job ($p < .05$). The work role-loss group did not exhibit lower professional role involvement or professional role identification, nor were they more likely to report work-related depression or lower life satisfaction than the nonreassigned group. In short, their discontent tended to focus on their new jobs but not on life in general or involvement and identification with their profession.

Losing one's job is a traumatic event (Finley & Lee, 1981; Hesson, 1978; O'Brien & Kabanoff, 1979). It involves the loss of a work role that contributes to the worker's status, identity, and feelings of self-worth. The loss of a job has a negative effect on the individual's attitudes toward the self, resulting in reduced self-esteem and a sense of lost identity (Hesson, 1978; Tabor, Walsh, & Cooke, 1976). A common reaction to unemployment is grief, of which anger and depression are components (Hesson, 1978; Schlossberg & Leibowitz, 1980).

The reactions to job loss are more than psychological, and frequently result in the onset of psychosomatic disorders or the flare up of old and formerly controlled ailments. A large-scale study of worker reactions to unemployment documented that a rise in the incidence of common colds and intestinal flu and increases in blood pressure, uric acid, serum cholesterol, and norepinephrine excretion occurred as a result of job loss (Cobb, 1974).

These studies of the responses to unemployment provide dramatic evidence of the negative effects of loss of one's job. However, in the study of the unemployed, the loss of the work role is confounded with the loss of income, benefits, employment, and involvement with the work organization. This article addresses the extent to which the negative effects of loss of one's job are due to the loss of a work role, apart from the loss of a livelihood and its concomitants of fringe benefits, work and social relationships, and associations with an employing organization.

The opportunity to separate work role-loss from employment loss was provided as a result of an unusual personnel decision by a large social services agency. In the administration of staffing cuts, a group of social workers were abruptly reassigned

to nonprofessional jobs. This natural manipulation provided an opportunity to study the impact of work role-loss, disassociated from employment and salary loss (the salary and benefits remained the same), on life satisfaction and work-related attitudes of job satisfaction, self-esteem, and depression, as well as involvement and identification with the professional role of social worker.

An understanding of the effects of work role-loss depends on a careful definition of several key concepts: job, work role, professional role, and the *fit* or consistency among them.

A job is usually taken to mean a set of tasks expected of occupants of specified positions. These tasks may be spelled out in a job description. *Work role* is a broader concept, defined as a set of expectations about behavior, attitudes, and values associated with a specified position. In much of the literature about work roles, the individual has been viewed as an interchangeable component in the organizational system, a passive role occupant (Graen, 1976). A more complex and realistic view suggests that the individual interacts with the role expectations, and may adapt, modify, or reject the pre-established role (Graen, Orris, & Johnson, 1973; Gutek & Morasch, 1982; Kahn, Wolfe, Quinn, Snoek, & Rosenthal, 1964; Ziller, 1964, 1965). For example, in their study of the assimilation of office workers into clerical jobs, Graen et al. (1973) found that about 40% fell into the role-rejecting category.

The findings of Graen et al. (1973) suggest the usefulness of another role concept: professional role. Professional role refers to the expectations about behavior and values associated with a particular occupational career or career path. Perhaps not everyone who is employed has a professional role, but for those who do, a professional role may or may not be consistent with the work role associated with a particular job. Inconsistencies between professional role and specific work role are important only when the individual has accepted as appropriate the behaviors and values of the professional role. When individuals accept and adopt the values and behaviors associated with the professional role, an inconsistency between professional role and work role translates directly into a poor fit between the person and the job. The concept of fit is important because research

We appreciate Dale Berger's statistical advice, Jim Cunningham's general advice and comments, and the superb suggestions of two anonymous reviewers whose recommendations greatly improved this article.

Correspondence concerning this article should be addressed to Barbara A. Gutek, Faculty in Psychology, Claremont Graduate School, Claremont, California 91711.

has shown that a good fit between person and job promotes beneficial outcomes such as job satisfaction, whereas a poor fit is associated with negative consequences (Brophy, 1959; Mount & Muchinsky, 1978). Many of the clerical workers studied by Graen et al. (1973) apparently experienced a poor fit between work role and professional role. Presumably, they could not modify the work role to fit better with their occupational identification. Thus, they were work-role rejectors.

Work roles involving professional qualifications generally are consistent with a professional role. Professionals such as social workers, lawyers, or managers spend many years in a job requiring specialized skills and knowledge. They are socialized into a professional role that they carry with them across different work roles.

The extent to which workers have adopted the values and behavior of a professional role is reflected in their involvement and identification with their occupation. Both involvement and identification are, unfortunately, "somewhat loose concepts" (Patchen, 1970, p. 152) that have been used in a variety of ways (Allport, 1947; Faunce, 1959; French, 1980; French & Kahn, 1962; Lawler, 1969; Lodahl & Kejner, 1965; Patchen, 1970; Rabinowitz & Hall, 1977). Here we are interested in professional role involvement and professional role identification.

The term *professional role involvement* is a measure of the importance of work in the individual's self-image (Lodahl & Kejner, 1965) and the extent to which one's occupation is important for life satisfaction (Lodahl, 1964). Professional role identification, on the other hand, is used to reflect the consciousness of belonging to a profession, career, or occupation (Kahn, 1981; Lawler & Hall, 1970; Maurer, 1969; Patchen, 1970). Professional role identification includes that definition of self that profession confers in terms of one's social status, values, and self-concept (Kahn, 1981; Lodahl & Kejner, 1965; Super, 1957; Super, Stanishevsky, Matlin & Jordon, 1963). Professional role identification is a standard by which a specific work role is evaluated as a good fit with one's future career plans and self-concept (see, e.g., Graen et al., 1973; Vroom, 1962).

If employees lose a work role that is compatible with a professional role by being placed in a job in which they cannot modify or alter the work role to be compatible with the expected behaviors of the professional role, they will experience a poor fit. The concept of *fit* seems useful to describe the relation between a set of professional behaviors and values and work role, although the concept has been used in the literature to describe the relation between personal characteristics and work role (Holland, 1973).

Under situationally induced circumstances such as a reassignment of employees to a different work role that results in a poor fit between work role and professional role, people are likely to exhibit negative effects. The expected effects are lowered job satisfaction, lowered self-esteem, and depression, the same effects exhibited by people who lose their job (which includes their work role; Hesson, 1978; Tabor et al., 1976).

We expect that the abrupt change in the work role experienced by the reassigned social workers results in a poor fit between the person's professional and work roles and will induce a stress reaction. Further, the reassigned workers will be placed in an environment in which their professional role identity as social workers is no longer reinforced or confirmed by their superiors, co-workers, and clients.

One way of coping with the poor fit that results from work role-loss would be to alter involvement and identification with one's professional role so that the new work role and occupational role are more compatible. Thus, workers may respond by lowering their identification and involvement with social work. On the other hand, given that social workers have invested heavily in their professional role, they may continue to be involved and identified with social work even though it constitutes a poor fit with their work role. They may choose to leave or intend to leave their new work role instead.

On the basis of the preceding discussion, we hypothesized that at the point of the survey, about 9 months following reassignment, the role-loss group, compared to the nonreassigned social workers, will show the following effects: (a) lower professional role involvement, (b) lower professional role identification, (c) lower self-esteem, (d) increased depression related to work, (e) lower job satisfaction, (f) lower general satisfaction, and (g) increased intention to leave the job.

Field Experiment

A large public social services agency on the west coast abruptly displaced approximately 180 social workers, creating a natural field experiment. In the administration of large staffing cuts, the agency made a decision to reassign "excess" personnel in social services to another bureau to process welfare applications and payments. These demotions resulted in the reassignment of employees with professional skills for social work to nonprofessional positions for which casework was not part of the duties and, in effect, was discouraged. The duties for the new clerical and administrative positions required interviewing applicants and verifying the required documentation to determine eligibility for assistance programs. This included completing the required forms to set up budget processes, as well as documenting that all government procedures and guidelines had been followed.

This natural manipulation offered the opportunity to study work role-loss that was disassociated from the losses of employment, salary, fringe benefits, and association with the organization. The reassignment did not result in a reduction of salary or fringe benefits.

Moreover, the procedure used for selection of the reassigned workers resulted in few differences between the role-loss and the nonreassigned groups. The reassignments were based on seniority within each job classification level; jobs were then reassigned within each district office. The excess personnel at each job classification level were reassigned to nonservices positions, rather than each classification being cascaded down to the next level in the job family. For example, all excess staff in the Social Worker III classification in each district office was reassigned to nonprofessional jobs instead of receiving a demotion to the next lower level services position of Social Worker II.

Method

Procedures

The research was carried out in two steps. First, 10 workers were interviewed in person during a pilot study the first week in November 1981, following the reassignment of the workers that began 1 month

previous to that time. The interviews were used to gain qualitative information to aid in the development of the survey questionnaire.

Second, a questionnaire was distributed between July 15 and August 1 of 1982 to the district offices, about 9 months following the reassignments. The interval of 9 months was judged to be sufficient time for workers to recover from the initial shock of reassignment and to adjust as much as they were likely to. A cover letter and consent form and a stamped, self-addressed envelope were provided with each questionnaire. The questionnaires were coded to protect the confidential nature of the information. All of the questionnaires were returned by August 30, 1982. A follow-up questionnaire was not possible because the agency was concerned about disruption of work.

Subjects

The subjects were 132 social workers. The reassigned workers (work role-loss group) were sampled from district offices from a personnel list of the 180 social workers who were reassigned. The questionnaires were distributed at the work locations and were mailed to workers who were out sick or had recently been transferred. In all, 66 reassigned workers responded out of a group of 90 workers who received the questionnaires.

The nonreassigned workers ($n = 66$) were sampled from five district offices. The sample included all social workers in the five district offices. In all, 66 subjects responded from a total of 134 workers who received the questionnaires. (The 49% response rate is viewed as adequate for analysis by Babbie, 1973.)

All of the subjects were college educated, and 42 had graduate degrees. The sample included 86 women and 46 men. The 132 subjects included 93 Whites, 18 Blacks, 13 Hispanics, and 8 Asians or others. The age range for the sample was from 26 to 66 years old, and the salary range was from \$18,000 to \$28,000 per year.

The only differences between the two groups were in age and tenure of the subjects with the organization. The role-loss group was about 4 years younger than the nonreassigned group (40.4 vs. 44.6 years old), $t(125) = -2.61, p < .01$, and had an average of about 2 years less tenure (11.1 vs. 13.3 years), $t(128) = -2.02, p < .05$. The distributions of sex, ethnicity, and marital status were not different for the two groups.

Measures

The questionnaire included indexes of work-related self-esteem, work-related depression, and a scale of facet-specific questions about satisfaction with the job context. A general, facet-free question about job satisfaction was added, and two questions to measure professional role involvement were included as well. The measure of work-related self-esteem and depression were taken from *The 1972-1973 Quality of Employment Survey* developed by the Survey Research Center at the University of Michigan (items are listed in Quinn & Shepard, 1974.) The facet-specific measure of job satisfaction was taken from *The 1977 Quality of Employment Survey* by Quinn and Staines (1979).

Work-related self-esteem. The five questions used to measure self-esteem on the job were each in the format of a 7-point bipolar scale (e.g., *successful* [7] vs. *not successful* [1]). This scale generated an alpha reliability coefficient for internal consistency of .77.

Depressed mood. A scale of 10 questions with 4-point response options was used to measure depression in the job-related context and generated an index with an internal consistency reliability of .86.

Job satisfaction. A facet-specific scale of seven questions regarding satisfaction with aspects of work was developed from a factor analysis of a job satisfaction scale used by the Survey Research Center in *The 1977 Quality of Employment Survey* (Quinn & Staines, 1979). The scale ($\alpha = .67$) consists of the following items, each having four response categories: "I am given a chance to do the things I do best," "My fringe benefits are good," "The physical surroundings are pleasant," "I am

free from the conflicting demands that other people make of me," "My supervisor is successful in getting people to work together," "Promotions are handled fairly," and "The people I work with take a personal interest in me."

A facet-free question regarding overall job satisfaction, with a 7-point scale ranging from *very satisfied* (7) to *very dissatisfied* (1), was also included. The item, "All in all, how satisfied are you with your job?" was also adapted from *The 1977 Quality of Employment Survey*.

Intention to leave the job. A question asking, "Taking everything into account, how likely is it that you will make an effort to find a new job with another employer within the next year?" was used to assess the intention to "turn over." The possible responses were *very likely*, *somewhat likely*, *somewhat unlikely*, and *not at all likely*.

Professional role involvement. Two questions were adapted from *The 1977 Quality of Employment Survey* (Quinn & Staines, 1979) to measure role involvement. "How much do you agree or disagree that the most important things that happen to you involve your job?" and "My main satisfaction in life comes from my work role as a social worker," measured on a 5-point scale, generated an index with an internal consistency of .70.

Professional role identification. Three individual questions were used to measure three aspects of the subject's identification with the role of social worker. An open-ended question, "If you were free to go into any type of job you wanted, what would you be doing 5 years from now?" was used. The responses were later categorized into two groups: those related to social work and those not related to social work. A question to measure the subject's view of career prospects asking, "How useful and valuable will your present social work skills be on the job market 5 years from now?" was used, with the response options on a 7-point bipolar scale of *very useful* (7) to *not at all useful* (1). A question to measure satisfaction with the role of social worker asked, "Knowing what you know now, if you had to decide all over again whether to take the job as a social worker, would you take it?" The response categories were *definitely yes*, *probably yes*, *probably not*, and *definitely not*.

Life satisfaction. Two questions assessing satisfaction and happiness, on a 7-point bipolar scale, were combined to form a scale of general satisfaction ($\alpha = .95$). The format was, "In general, how satisfying do you find your *present life*? Circle the number that best describes how you see your life (happy vs. unhappy and satisfying vs. unsatisfying)."

Results

Summary of Pilot Study Interviews

A major focus of the 10 interviews was to obtain qualitative information regarding the effect of the loss of work role on attitudes about the self and to assess the degree and number of physiological symptoms.

All of the subjects indicated a sense of loss with some loss of self-esteem. There was a range in the number and degree of physiological symptoms. The range of symptoms per subject (headaches, flu, colds, vomiting, diarrhea, stomach pains and nausea, and flare-ups of controlled ailments) was from one to five; 3 of the 10 subjects required treatment by a physician. In all, 7 of the 10 subjects became physically ill within 2 days of reporting to the nonprofessional assignment, and missed 2 or more days of work the first week on the new assignment. A total of 8 workers said they were mildly to moderately depressed.

The actual statements of 3 subjects provide a rich illustration of these personal reactions. The first subject is a woman, 52 years old, married and a parent, having worked 14 years with the agency.

I guess I am mourning a loss of my friends, all that is familiar. A

Table 1
Mean Responses of Role-Loss Group and of the Nonreassigned Group to Measures of Dependent Variables

Dependent variable	M		t test significance
	Role-loss group	Nonreassigned group	
Professional role involvement	2.9	3.1	−0.91
Professional role identification			
Take job again	2.9	2.8	0.26
Social worker skills useful	4.0	3.5	1.48
Self-esteem	4.1	4.9	−3.52**
Facet-free job satisfaction	2.5	4.1	−5.31**
Facet-specific job satisfaction	2.1	2.3	−2.10*
Life satisfaction	4.9	5.4	−2.01*
Intention to leave current position	2.8	2.2	3.05**
Work-related depression	2.6	2.8	−1.62

Note. Larger values indicate higher amount of variables, except for depression in which lower score indicates greater depression.
* $p < .05$. ** $p < .01$.

loss of power to make important decisions. It brings up old feelings of inadequacy. Triggers old rejections—after all [laughs], mom always liked my brothers best. After 3 days of that idiotic training (for new positions) I was nearly crazy with rage. . . . Then I got hold of myself. . . . But afterwards I was very sick.

The second subject, a woman, 40 years old and single, worked 13 years with the agency.

They called me into the director’s office and told me I had 1 week before I was to report to eligibility. . . . Stunned, I could not believe it. The following week I told myself they would call it off at the last minute. That Friday before reporting, I went home ill: vomiting, diarrhea, upset stomach. I was sick for 3 days but I made myself go (to work). . . . It’s a loss of status, your identity is tied up in what you do. A loss of my chosen career. . . . I cannot believe they would do that to a valued player.

The third subject is a man, 66 years old, married and a parent, having worked 16 years with the agency.

I feel diminished in my role and personality. I am somewhat despondent. It involves a loss of personal dignity, a loss of personal standing. I have 16 years with the agency and 12 years as a services supervisor. . . . My job was important—working with adults who were abandoned and exploited. I am a social worker not a paper pusher.

Results of Survey

The first issue to be examined is whether people who lost their work roles also changed their professional role involvement and professional role identity in the 9 months following reassignment. It was expected that if the work role-loss engendered change in professional role identity and role involvement, those social workers who had lost their work roles would show lower identification and involvement than would the nonreassigned workers.

Table 1 shows that the expected differences were not found: In regard to the questions about the subject’s professional role involvement as a social worker, the responses of the role-loss group were similar to those of the workers who retained their work roles. Two questions assessing professional role identification—whether the worker would take the job as a social worker

again, and how useful and valuable the social work skills were perceived to be in 5 years—did not reveal any significant differences between the two groups. A chi-square analysis of the responses to the third role identification question about the subject’s job aspirations in 5 years did not yield any differences between the groups, $\chi^2(1, N = 132) = .00$.

The second issue to be addressed is whether work role-loss is associated with lower levels of self-esteem, facet-free job satisfaction, facet-specific job satisfaction, life satisfaction, the intention of remaining on one’s job, and increased work-related depression.

Work role-loss was related to all of these measures in the expected direction, although the difference did not reach statistical significance for work-related depression (see Table 1). Work role-loss was, however, associated with lower levels of self-esteem, job satisfaction, and life satisfaction, and higher self-reported intention to turn over.

Not surprisingly, work role-loss was more strongly related to facet-free job satisfaction than to facet-specific job satisfaction ($p < .001$ vs. $p < .05$). The majority of the items in the facet-specific measure did not reflect significant differences between reassigned and nonreassigned workers. A disaggregation of the seven items in the facet-specific measure revealed that only two items differentiated the groups. The reassigned group was significantly lower, $t(129) = -6.51, p < .001$, on the items “I am given a chance to do the things I do best” and “My supervisor is successful in getting people to work together,” $t(129) = -2.02, = p < .05$. There were no differences on the other five items.

It is possible that the differences outlined in the work-related attitudes between the two groups may be artifactual, arising from the small but significant demographic differences between the two groups of workers on the variables of age and tenure. In addition, there may be unanticipated effects contributed by the other demographic variables of sex, ethnicity, and marital status.

To gather information about the possibility of such effects, each of the dependent variables shown in Table 1, for which the role-loss group and the nonreassigned group differed (self-

Table 2

Demographic Characteristics and Role Loss as Predictors of Work-Related Self-Esteem, Facet-Free Job Satisfaction, and Intention to Leave Job

Predictor variable	Self-esteem		Job satisfaction		Intention to leave	
	Final β	R^2 added	Final β	R^2 added	Final β	R^2 added
Age	.04	.01	.07	.03	-.04	.07**
Sex ^a	-.10	.01	-.18*	.03*	.09	.01
Ethnicity ^a	.02	.00	-.01	.01	.00	.00
Marital status ^a	-.10	.02	-.13	.03	.06	.01
Tenure	.02	.00	.02	.00	-.37**	.10**
Role loss ^a	.27*	.06**	.38**	.12**	-.17*	.03*
Cumulative R^2	.11		.22		.22	

Note. $N = 132$.

^a Sex, ethnicity, marital status, and role loss were dummy coded, with the larger value associated with female, Caucasian, married, and nonreassigned, respectively.

* $p < .05$. ** $p < .01$.

esteem, facet-free job satisfaction, facet-specific satisfaction, life satisfaction, and intention to leave job), was regressed on the demographic characteristics and the presence or absence of work role-loss. These regression analyses indicated that when differences in the demographic characteristics are statistically controlled, role loss remains a significant predictor for three of the five dependent variables. Specifically, role loss was associated with lower self-esteem (Table 2, $r = .30$, $p < .001$, R^2 added = .06, $p < .01$), and lower facet-free job satisfaction (Table 2, $r = .42$, $p < .001$, R^2 added = .12, $p < .001$). In the case of facet-free satisfaction, sex was also a significant but much weaker predictor ($r = -.19$, $p < .05$, R^2 added = .03, $p < .05$), with women showing lower facet-free satisfaction than men. In addition, role loss was associated with lower intention to remain on the job (Table 2, $r = -.26$, $p < .05$, R^2 added = .03, $p < .05$). In the case of intention to leave, tenure was also a highly significant predictor (Table 2, $r = -.42$, $p < .001$, R^2 added = .10, $p < .001$), with subjects lower in tenure expressing greater willingness to find another job. With the demographic variables controlled, facet-specific job satisfaction and life satisfaction were lower for the role-loss group, but did not reach significance.

The preceding findings suggest that the covariance between role loss and the dependent variables of self-esteem, overall job satisfaction, and intention to remain on the job cannot be explained by differences in age, sex, ethnicity, marital status, or tenure.

Discussion

This natural experiment provided an opportunity to study the impact of work role-loss among a group of professionals while controlling the effects of the context factors of salary, fringe benefits, and work environment. Responses of the reassigned workers interviewed during the pilot study (about 1 month after reassignment) illustrated that the abrupt work role-loss induced a crisis reaction similar to that described in the literature about the reactions of the recently unemployed (cf. Finley & Lee, 1981; Tabor et al., 1976). The 10 subjects de-

scribed feelings of lost identity, loss of status, reduced feelings of self-worth, and increased incidence of illness and physiological complaints.

The survey results obtained 9 months after the reassignment suggest that the work role-loss group had adapted to the new role inasmuch as the emotional distress, as revealed by their responses, was minimal. Their mean scores on the measure of work-related depression were not significantly different from the nonreassigned group. It is possible that the new jobs were less stressful and less demanding and that these lower demands offset some of the effects of their professional role loss. Scores on life satisfaction were not significantly lower for the reassigned group when various demographic characteristics were controlled. Even if reassigned workers had adjusted to their new jobs, they did not value them. Although the reassigned social workers exhibited minimal emotional distress, they reported lower job satisfaction and greater intentions to turn over than did the nonreassigned workers.

In addition, the survey results indicated that the role-loss group had not lowered their professional work involvement or identification to fit their new jobs. The interviews with the workers made at the time of the survey suggest high involvement and identification with the role of social worker and a rejection of the new role, supporting Graen et al.'s (1973) definition of the "role rejector." One individual said that he could learn to do the new job well, but that he did not have any interest in doing so because he did not plan to continue the new job any longer than was necessary to locate work in his occupation. Several of the workers in the role-loss group remarked that they were planning to return to graduate school.

The importance and persistence of professional role involvement and identification are suggested by these findings. In a professional occupation that entails strong role identification and role involvement with that profession, identification and involvement are likely to remain, even though a compatible work role is lost. Social workers, like other professionals, spend years preparing for and learning their professional roles. The identification and involvement that develops does not simply

disappear if a compatible work role disappears. Consequently, it is not surprising that under circumstances of work role-loss, social workers reported lowered job satisfaction and self-esteem and a heightened interest in finding another job.

These data support Patchen's (1970) hypothesis that when employees are identified and involved with their professional roles, effective performance of those valued activities may engender satisfaction with the work. The results of this study illustrate the importance of the fit between work role and the valued activities of one's profession in contributing to the quality of working life. The findings lend support for Holland's (1973) contention that if there is a good fit between people's interests and the activities of their work role, they are more satisfied with their job and have longer tenure.

Work role-loss does not necessarily spill over to other areas of life, contrary to Holland's (1973) theory. The role-loss group confined their dissatisfactions to their new jobs. Statements of several of the workers from the work role-loss group suggested that they coped with the loss of work role by adjusting their expectations about the contribution their job would make to their life satisfaction. A statement of one worker illustrates this adjustment in values:

I devote much less time to the job and give more time and importance to other areas of my life. . . . I notice I concentrate less on what people do for a living and more on what they are as people.

These results point to a distinction between professional role involvement and identification and job involvement and identification. The social workers did not evince less identification with social work than did the nonreassigned. However, when the social work role was lost, they did indicate lower involvement with the new job and increased intention to find another job. The results suggest an alternative perspective of professional role identification and role involvement as mediating variables in the research on work-related attitudes.

Finally, the differential scores on the two job satisfaction measures support the validity of the measures.¹ Reassigned and nonreassigned groups showed a greater difference on facet-free than on facet-specific job satisfaction. This is not surprising given that facet-specific satisfaction assesses aspects of the job that were the same for both groups (i.e., fringe benefits, physical surroundings), as well as items on which the two groups are likely to differ substantially (e.g., "I am given a chance to do the things I do best"). The facet-free measure is less specific and presumably reflects whatever parts of the job are most important to respondents. Hence, facet-free satisfaction reflects all aspects of the job, taking into account (weighted by) individuals' perceptions of the importance of each component (Quinn & Mangione, 1973). As such, facet-free satisfaction is a more sensitive barometer for assessing change in important areas like work role when not all aspects of the job assessed in facet-specific measures (e.g., salary or fringe benefits) have changed.

In summary, the strength of our findings resides in the opportunity to assess the reactions of the subjects to an unusual natural experiment in which work role-loss occurred without the concomitant losses of the extrinsic rewards of employment, salary, and fringe benefits. Survey data, assessing the impact of work role-loss after a lapse of about 9 months, suggest that the work role-loss group adapted to the poor person-role fit over

time in that emotional distress was minimal. Job satisfaction, however, was low for the role-loss group, and they reported lower self-esteem and a higher intention to leave the job than did the nonreassigned group.

Although this study was limited to one professional group, other studies also support the importance of fit in contributing to job satisfaction (Brophy, 1959; Caplan, Cobb, French, Van Harrison, & Pinneau, 1975; Dore & Meacham, 1973; French, Rodgers, & Cobb, 1974). The issue of whether the consequences of poor fit are equally powerful in all occupational groups, however, needs to be addressed. These findings that show a strong relation between work role-loss and lowered job satisfaction, lowered self-esteem, and greater intention to turn over may hold only for occupations in which strong professional role involvement and identification are evinced.

¹ This point was called to our attention by one of the anonymous reviewers.

References

- Allport, G. W. (1947). The psychology of participation. *Psychological Review*, 52, 117-132.
- Babbie, E. R. (1973). *Survey research methods*. Belmont, CA: Wadsworth.
- Brophy, A. L. (1959). Self, role and satisfaction. *Genetic Psychological Monograph*, 59, 263-308.
- Caplan, R. D., Cobb, S., French, J. R. P., Jr., Van Harrison, R., & Pinneau, S. R., Jr. (1975). *Job demands and worker health* (Research report). Cincinnati, Ohio: National Institute of Occupational Safety and Health.
- Cobb, S. (1974). Physiologic changes in men whose jobs were abolished. *Journal of Psychosomatic Research*, 18, 245-258.
- Dore, R., & Meacham, M. (1973). Self-concept and interests related to job satisfaction of managers. *Personnel Psychology*, 26, 49-59.
- Faunce, W. (1959, August). *Occupational involvement and the selective testing of self-esteem*. Paper presented at the meeting of the American Sociological Association, Chicago.
- Finley, M. H., & Lee, A. T. (1981). The terminated executive: It's like dying. *Personnel and Guidance Journal*, 59, 382-384.
- French, J. R. P., Jr. (1980). Person-role fit. In D. Katz, R. L. Kahn, & J. S. Adams (Eds.), *The study of organizations* (pp. 444-450). San Francisco: Jossey-Bass.
- French, J. R. P., Jr., & Kahn, R. (1962). A programmatic approach to studying the industrial environment and mental health. *Journal of Social Issues*, 18, 1-47.
- French, J. R. P., Jr., Rodgers, R., & Cobb, S. (1974). Adjustment as person-environment fit. In G. V. Coelho, D. A. Hamburg, & J. E. Adams (Eds.), *Coping and adaptation* (pp. 316-333). New York: Basic Books.
- Graen, G. (1976). Role-making processes in organizations. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 1201-1243). Chicago: Rand McNally.
- Graen, G., Orris, J. B., & Johnson, T. (1973). Role assimilation processes in a complex organization. *Journal of Vocational Behavior*, 3, 395-420.
- Guterk, B. A., & Morasch, B. (1982). Sex-ratios, sex-role spillover, and sexual harassment of women at work. *Journal of Social Issues*, 38, 55-74.
- Hesson, J. E. (1978). The hidden psychological costs of unemployment. *Intellect*, 106, 389-390.
- Holland, J. L. (1973). *Making vocational choices: A theory of careers*. Englewood Cliffs, NJ: Prentice-Hall.

- Kahn, R. L. (1981). *Work and health*. New York: Wiley.
- Kahn, R. L., Wolfe, D. M., Quinn, R. P., Snoek, J. D., & Rosenthal, R. A. (1964). *Organizational stress: Studies in role conflict and ambiguity*. New York: Wiley.
- Lawler, E. E., III. (1969). Job design and employee motivation. *Personnel Psychology*, 22, 426-435.
- Lawler, E. E., III, & Hall, D. T. (1970). Relationship of job characteristics to job involvement, satisfaction, and intrinsic motivation. *Journal of Applied Psychology*, 54, 305-312.
- Lodahl, T. M. (1964). Patterns of job attitudes in two assembly technologies. *Administrative Science Quarterly*, 8, 482-519.
- Lodahl, T. M., & Kejner, M. (1965). The definition and measurement of job involvement. *Journal of Applied Psychology*, 49, 24-33.
- Maurer, J. G. (1969). *Work role involvement of industrial supervisors*. East Lansing, MI: MSU Business Studies.
- Mount, M. K., & Muchinsky, P. M. (1978). Person-environment congruence and employee job satisfaction: A test of Holland's theory. *Journal of Vocational Behavior*, 13, 84-100.
- O'Brien, G. E., & Kabanoff, B. (1979). Comparison of unemployed and employed workers on work values, locus of control and health variables. *Australian Psychologist*, 14, 143-155.
- Patchen, M. (1970). *Participation, achievement and involvement on the job*. Englewood Cliffs, NJ: Prentice-Hall.
- Quinn, R. P., & Mangione, T. (1973). Evaluation weighted models of job satisfaction: A cinderella story. *Organizational Behavior and Human Performance*, 10, 1-23.
- Quinn, R. P., & Shepard, L. T. (1974). *The 1972-1973 Quality of Employment Survey: Descriptive statistics with comparison data from the 1969-70 survey of working conditions*. Ann Arbor: University of Michigan Survey Research Center, Institute of Social Research.
- Quinn, R. P., & Staines, G. L. (1979). *The 1977 Quality of Employment Survey: Descriptive statistics with comparison data from the 1969-70 and the 1972-73 surveys*. Ann Arbor: University of Michigan Survey Research Center, Institute for Social Research.
- Rabinowitz, S., & Hall, D. T. (1977). Organizational research on job involvement. *Psychological Bulletin*, 84, 265-288.
- Schlossberg, N. K., & Leibowitz, Z. (1980). Organizational support systems as buffers to job loss. *Journal of Vocational Behavior*, 17, 204-217.
- Super, D. E. (1957). *The psychology of careers*. New York: Harper & Row.
- Super, D. E., Stanishevsky, R., Matlin, N., & Jordon, J. P. (1963). *Career development: Self-concept theory*. Princeton, NJ: College Entrance Examination Board.
- Tabor, T. D., Walsh, J. T., & Cooke, R. A. (1976). *Report on an innovative corporate/community program for reducing the social impact of a plant closing*. Ann Arbor: University of Michigan, Institute for Social Research.
- Vroom, V. H. (1962). Ego-involvement, job satisfaction, and job performance. *Personnel Psychology*, 15, 159-177.
- Ziller, R. C. (1964). Individuation and socialization: A theory of assimilation in large organizations. *Human Relations*, 17, 341-360.
- Ziller, R. C. (1965). Toward a theory of open and closed groups. *Psychological Bulletin*, 64, 164-182.

Received July 20, 1986

Accepted September 13, 1986 ■

The Social Psychology of Eyewitness Accuracy: Misleading Questions and Communicator Expertise

Vicki L. Smith and Phoebe C. Ellsworth
Stanford University

In two studies we examined the effect of questioner expertise on the error rates of subjects who were asked misleading versus unbiased questions. A total of 105 introductory psychology students watched a videotaped clip of a bank robbery and were then questioned about the crime. The questioner was represented to subjects as either highly knowledgeable or completely naive about the events the subject witnessed. One half of the subjects in each expertise condition were asked misleading questions, and the other half were asked unbiased questions. In the knowledgeable questioner conditions, misleading questions were associated with error rates significantly higher than those obtained with the unbiased questions ($p < .05$). In the naive questioner conditions, equivalent error rates for both types of questions were obtained (ns). These results indicate that misleading questions decrease witness accuracy when the questioner is assumed to be knowledgeable about the crime, but have no effect on accuracy when the questioner is assumed to be naive.

Legal scholars and experimental psychologists share a long-standing concern about the factors that make an eyewitness's testimony accurate and, perhaps more common in the experimental literature, the factors that impair accuracy. An important general issue raised in this work is the malleability of an eyewitness's memory. Is it possible to alter a person's memory for a crime, and if so, how does this process operate?

The study of misleading questions and their effects on accuracy is the approach most commonly taken by recent researchers who have worked on the question of memory malleability. Misleading questions provide information that is inconsistent with the event witnessed, suggesting, for example, the existence of an object that was in fact not present. Elizabeth Loftus and her colleagues (Loftus, 1975, 1977; Loftus, Altman, & Geballe, 1975; Loftus, Miller, & Burns, 1978; Loftus & Palmer, 1974; Loftus & Zanni, 1975) have repeatedly demonstrated that subjects who are asked misleading questions tend during subsequent questioning, to report the false information contained in the misleading questions they were asked. As a result, these subjects make significantly more errors in their reports of the event than do subjects who are asked unbiased questions. For example, in one experiment (Loftus, 1975), subjects watched a videotape of a traffic accident and then answered questions about the incident. One half of the subjects were asked, "How fast was the

white sports car going when it passed the barn while traveling along the country road?" This question is misleading because it contains the presupposition that there was a barn in the clip, when in fact no barn was shown. The remaining subjects were simply asked, "How fast was the white sports car going while traveling along the country road?" When later asked if they had seen a barn, 17.3% of the subjects exposed to the misleading question reported having seen a barn, whereas only 2.7% of the control subjects made this error.

Loftus (1975, 1977) explained this phenomenon by proposing that the false presupposition contained in the misleading question is incorporated into the witness's memory for the event. When asked to recall the incident, the subject remembers a barn but fails to recall that the source of this memory was the question rather than the event itself. As a result, the information is recalled and reported as part of the event that the subject witnessed. Using this same general paradigm, Loftus and others (Clifford & Scott, 1978; Lesgold & Petrush (cited in Loftus, 1979); Loftus, 1977; Loftus et al., 1975; Loftus et al., 1978; Loftus & Palmer, 1974; Loftus & Zanni, 1975) have replicated these results with a variety of stimulus materials and have obtained error rates as high as 95% for subjects who are asked misleading questions (Clifford & Scott, 1978).

Critics have argued that this effect of misleading questions is actually less general than these high error rates imply. Dritsas and Hamilton (cited in Loftus, 1979) reported that memory for items that are salient and central to the witnessed event is significantly less likely to be altered by misleading information than is memory for peripheral details. These authors maintained that the high error rates reported by other investigators are attributable to the testing of peripheral details.

Yuille (1980) found that misleading information did not affect the accuracy of subjects whose memories already contained a correct representation of the implied object. Subjects viewed slides that depicted an accident involving an automobile and a pedestrian, and then described the event in writing. Following this free-recall stage, subjects completed a questionnaire

This research was conducted while the first author was supported by a National Science Foundation Graduate Fellowship.

The authors are grateful to Robert Mauro for his help in various phases of this research, and to R. Edward Geiselman for providing the crime clip used in Experiment 2. We would also like to thank Pamela Burke, Craig Smith, and Anthony Ahrens for their helpful comments on an earlier version of this article, and Andrew Blaine, Gary Rothchild, Kazuo Sano, and Michael Tuchin for serving as confederates in these experiments.

Correspondence concerning this article should be addressed to Vicki Smith, Department of Psychology, Stanford University, Stanford, California 94305.

about the accident that either did or did not contain misleading information about the type of traffic sign present at the intersection. Overall, the misleading question effect was replicated: Subjects who were asked misleading questions were significantly more likely to be wrong about the traffic sign than were subjects who were asked unbiased questions. However, only subjects who failed to mention the sign in their original reports of the event were misled by the false information. Subjects who correctly described the type of traffic sign during the free-recall stage were relatively immune to the false information.

Together, the results of these two experiments indicate that uncertainty is an important antecedent to the misleading question effect. When the subject's memory already contains a clear representation of the object, false information suggested afterward is not incorporated into memory. As proposed by Dritsas and Hamilton (cited in Loftus, 1979), the high error rates obtained by some investigators may be a function of their focus on tangential details that subjects did not carefully encode in memory.

So far, both the original researchers and their critics have considered the power of misleading questions to distort memory as a purely cognitive phenomenon. The vast social psychological literature on persuasion and attitude change (cf. McGuire, 1969) has not been brought to bear on the interactions between a questioner and a witness. In this article, we examine the effects on eyewitness accuracy of one of the classic social psychological variables—the expertise of the communicator (cf. Hovland & Weiss, 1951). Others have suggested that the credibility of the questioner may influence the accuracy of a witness's report (e.g., Yuille, 1980), but no systematic study of this issue has been undertaken.

Misleading information is typically presented to subjects on a questionnaire designed by the experimenter. Having designed the original event, the experimenter is necessarily well acquainted with it, and is therefore a highly credible source of information. The subject, having seen the event only once, assumes that he or she knows less about the details of the event than does the experimenter, and thus is likely to accept the information suggested by the experimenter without questioning its accuracy. The subject assumes that the new information is correct, and modifies his or her memory of the event to make it consistent with that information.

Suppose, however, that the source of the misleading information were less credible than the experimenter. When interrogated by someone known to be unfamiliar with the witnessed event, the subject would be less likely to accept the misleading information as accurate and therefore less likely to modify his or her memory of the event. Under these circumstances, asking misleading questions should not diminish the witness's accuracy. Thus, the influence of misleading questions may well depend on the witness's assumptions about how much the questioner already knows.

Investigation of the effects of the questioner's knowledge of the crime also has practical implications for the questioning of real-world witnesses. Experiments on eyewitness questioning have generally examined the effects of misleading information presented during the initial questioning of the subject. This situation is analogous to the questioning of real-world witnesses by police at the scene of the crime. Because the police are typically called to the scene after the crime has been committed,

they don't know what actually happened, but must try to reconstruct the event from the witness's report. Although there are some situations in which a police questioner already knows (or thinks he or she knows) a great deal about the crime, the initial questioning of a witness is often conducted by naive police questioners. Thus, it is important to know whether communicator expertise is a significant moderating variable affecting the power of misleading questions.

We designed two experiments to compare the relative impact on accuracy of different levels of questioner expertise by manipulating the knowledge of the questioner. Consistent with previous research, one experimental condition portrayed the questioner as being well acquainted with the crime that the subjects witnessed. In the other experimental condition, the questioner was represented to subjects as being unaware of what had taken place. We predicted that when the questioner was seen as knowledgeable about the crime, the misleading question effect obtained in previous research would be replicated. However, when the questioner was represented as being naive to the witnessed event, we expected that the misleading question effect would disappear and that subjects in this condition would make no more errors than did control subjects.

Experiment 1

Method

Overview

Subjects watched a videotaped clip of a bank robbery, after which a confederate of the experimenter asked the subject a series of questions about the crime. Two variables related to this questioning were factorially crossed. The first was the questioner's knowledge of the crime: Subjects were interrogated by a questioner who was either knowledgeable or naive about the crime the subject witnessed. The second factor manipulated was the type of question asked: One half of the subjects were asked misleading questions, and the other half were asked unbiased questions. After a 20-min filler activity, subjects completed a questionnaire about the crime that included critical questions designed to assess the effects of the misleading information.

Subjects

A total of 45 undergraduates enrolled in introductory psychology participated individually in the experiment, and were randomly assigned to one of the four experimental conditions. Fifteen subjects participated in each of the two misleading question conditions; the remaining 15 subjects were assigned to the unbiased question groups. Because no misleading information was given to subjects in the unbiased question conditions, it was expected that equivalent error rates would be obtained for these two groups, permitting their combination into a single control group with 15 subjects.

The data for 6 subjects were excluded from the experiment, with comparable subject loss in each condition. Three subjects were dropped because they did not give yes or no responses to the critical items on the crime questionnaire, but wrote ambiguous answers. Another was excluded because he recognized the film clip that was used and had some difficulty remembering what information he knew from the clip and what he remembered from the remainder of the film. Another was dropped because he was mistakenly asked the misleading questions while being assigned to the control condition. Finally, 1 subject was deaf, and thus unable to either process the auditory portion of the clip or understand the questions he was asked. Six additional subjects (ran-

domly assigned) replaced those excluded so that the total number of subjects participating in the experiment remained at 45.

Two undergraduate men served as confederates. The confederate's task was to portray himself as another introductory psychology student taking part in the experiment for course credit. The confederates were trained to assume this role convincingly, with as much cross-confederate behavioral consistency as possible.

Procedure

On arriving at the laboratory, the subject was directed to one of two chairs positioned in front of a videotape monitor and was told that the experiment would begin as soon as the other subject arrived. The confederate always arrived a few minutes later than the subject did to avoid suspicion by subjects of experimenter–confederate collusion. The experimenter began the session by explaining that the study dealt with the questioning of eyewitnesses to criminal events and, particularly, with how complete and accurate a report of a crime someone could get by questioning an eyewitness.

The experimenter then went on to explain that one of the participants would assume the role of eyewitness and the other the role of questioner. The procedure for assigning the subject the role of eyewitness varied with the experimental condition. Subjects in the naive questioner conditions drew lots, arranged so that the subject was always the witness. In the knowledgeable questioner conditions, the experimenter explained that the role assignments had been randomly determined prior to the session, and identified the subject as eyewitness and the confederate as questioner. The participants were then told that the questioner would move to another room while the witness watched a videotaped clip of a crime, and that the questioner should use this time to think of questions to ask the witness that would reveal as much detailed information about the crime as possible.

Questioner's knowledge. Critical to the hypotheses being tested was the manipulation of the questioner's knowledge of the crime. To this point in the procedure the subject was led to believe that the confederate was merely another subject in the experiment and thus equally unfamiliar with the videotaped clip. For subjects in the naive questioner conditions, this assumption was preserved and the questioner was merely told, "To make your job (of questioning) a little easier, I'll tell you that the crime (subject) will witness is a bank robbery." For subjects in the knowledgeable questioner conditions, the confederate was portrayed as well acquainted with the crime. To make this salient to the subjects, the following interchange took place:

Experimenter (to subject): To make (confederate's) job a little easier, I called and asked him to come in earlier today to familiarize himself with the videotaped clip you'll be seeing. He was given as much time as he needed to watch the clip over and over again and get a good idea of what happens, so that when the time came he could ask you good questions. He wasn't at that time asked to write down any questions; that's what he gets to do now.

Experimenter (to confederate): How many times did you end up seeing the clip?

Confederate: Oh, I don't know, 9 or 10 I guess.

Experimenter (laughing): Nine or 10? You probably know it better than I do.

Finally, the experimenter explained that later in the hour the subjects would be asked to fill out a questionnaire about the crime to see how effective the questioning process was. At this point, the confederate left the room and the subject watched the videotaped crime. The crime was a bank robbery that lasted just over 1 min.

Misleading and unbiased questions. After the subject had seen the crime, the confederate returned to the room with a list of 19 questions for the witness. These questions included the misleading information manipulation. For subjects in the misleading question conditions, 5 of

Table 1
Interrogator's Questions in Experiment 1

No.	Question
1.	How many robbers were there?
2.	What did they look like?
3.	Were they wearing masks?
4.	Were they wearing jackets? What kind? Color?
5.	What kind of pants were they wearing?
*6.	What kind of gloves were they wearing? (Were they wearing gloves?)
7.	What kind of weapons did the robbers have?
8.	Describe the gun the blue guy had. Was it small, large? Was it a pistol?
*9.	What did the other guy's gun look like? (Did the other guy have a gun?) Was it small, large? Was it a pistol?
10.	Did the robbers get away?
11.	Were there any police there? How many?
12.	How many tellers were there?
*13.	Not counting the tellers or the bank manager, how many witnesses were there? (Not counting the tellers, how many witnesses were there?)
*14.	Where was the getaway car parked? (Was there a getaway car?)
15.	What did the robbers say?
*16.	How many shots were fired? (Were any shots fired?)
17.	What time of day was the robbery?
18.	Was anybody tied up?
19.	Did everybody in the bank realize what was going on?

Note. Asterisk denotes questions containing misleading information; the unbiased versions of the questions appear in parentheses.

the questions contained misleading information; for subjects in the unbiased question conditions, the misleading information was removed. Table 1 lists the questions asked of subjects in the misleading question conditions, with the unbiased versions of the questions in parentheses. Because these questions were allegedly composed by the confederate during the time the subject watched the videotaped crime clip, they had to be carefully worded so as to arouse no suspicion by subjects of experimenter–questioner collusion.

Limitations on the content of the questions were imposed by the expertise manipulation. Because the same set of questions was asked by both knowledgeable and naive questioners, the topics covered had to be sufficiently general that the questions could logically be asked by both types of questioners. Thus, very specific questions about the robbery—such as, "Which police officer entered through the back of the bank?"—were avoided because they included details that could not possibly be known by a questioner who had never seen the crime clip. Note that in much of the previous research on this topic, the questions themselves imply expertise, perhaps further strengthening the subjects' impression that the experimenter must be knowledgeable. For example, it is highly unlikely that someone who knew nothing about an event would ever ask a question like, "How fast was the white sports car going when it passed the barn while traveling along the country road?"

The experimenter told the confederate that he would have 5 min to question the witness, stressed the importance of asking specific questions, and asked him to write down the witness's responses. The experimenter left the room during the interrogation, but returned at a predetermined point in the questioning process. Then, while the experimenter allegedly scored the questions and answers, the subject and the confederate completed a 20-min filler activity, which consisted of two questionnaires unrelated to the experiment.

Dependent measure. In the final stage of the experiment, subjects completed a 35-item questionnaire about the crime. This questionnaire

Table 2
Experiment 1: Percentage of Subjects Making Errors

Questioner type	Type of question	
	Misleading	Unbiased
Knowledgeable	73 (<i>n</i> = 15)	43 (<i>n</i> = 7)
Naive	27 (<i>n</i> = 15)	25 (<i>n</i> = 8)

was identical for subjects in all four conditions and contained 5 items tapping the subject's memory for the critical details: Were the robbers wearing gloves? Did you see a bank manager? Were any shots fired? Did both of the robbers have guns? Did you see a getaway car? Each of these questions required the subject to make a yes or no response, and an error was recorded when the subject responded yes to one or more of the items.

Results

Percentages of subjects who made errors in each of the four conditions are shown in Table 2. A *z* test on the arcsine transformed proportions (Langer & Abelson, 1972; Mosteller & Tukey, 1949) of subjects making errors in the two unbiased question groups revealed, as predicted, no significant difference in error rates for subjects questioned by a naive questioner and those questioned by a knowledgeable questioner, $z(\infty) = 0.74$, *ns*. Because the error rates obtained for the two unbiased question conditions did not differ significantly, these groups were combined for all subsequent analyses.

Two hypotheses examining the role of the questioner's level of expertise in producing the misleading question effect were tested. Specifically, it was hypothesized that (a) when the questioner was portrayed as more knowledgeable about the crime than were the subjects themselves, the misleading question effect obtained in previous research would be replicated, and (b) when the questioner was represented as naive to the crime, the misleading question effect would not be obtained and subjects in this condition would exhibit an error rate equivalent to that obtained in the unbiased question conditions.

Contrasts testing these two hypotheses were performed on the arcsine transformed proportions of subjects making errors in each condition (collapsed across the two unbiased question groups). As predicted, subjects who were asked misleading questions by a knowledgeable questioner made significantly more errors than did either subjects who were asked misleading questions by a naive questioner or subjects who were asked unbiased questions, $F(1, \infty) = 7.89$, $p < .01$. The second hypothesis was also confirmed: Misleading questions asked by a naive questioner produced no more errors than did unbiased questions, $F(1, \infty) = 0.13$, *ns*. Thus, the misleading question effect was obtained only in the condition in which the questioner was more knowledgeable about the crime than was the witness.

Examination of the data for each misleading question revealed that one of the questions accounted for the majority of the overall effect obtained in the knowledgeable questioner condition. The question, "What did the other guy's gun look like?" exhibited the highest error rate of the five misleading questions.

A contrast performed on the data for this question alone revealed that consistent with the overall effect, subjects who were asked misleading questions by a knowledgeable questioner made significantly more errors than did subjects in any other group, $F(1, \infty) = 9.85$, $p < .01$. The remaining four critical questions were associated with very low error rates in all conditions, precluding meaningful contrast analyses of these questions.

Separate chi-square tests were performed on the data for each critical question. Because our overall contrast analysis revealed no significant differences in the proportion of subjects who made errors when asked either misleading questions by a naive questioner or unbiased questions, these three conditions were combined for the chi-square analyses. As a result, our chi-square tests compared the frequency of error among subjects who were asked misleading questions by a knowledgeable questioner to the other three conditions combined. The question, "What did the other guy's gun look like?" yielded a highly significant chi-square value, $\chi^2(1, N = 45) = 10.15$, $p < .01$. Slightly more than one half of the subjects who were asked this question by a knowledgeable questioner later "remembered" seeing the nonexistent gun. Only 3 of the 30 subjects in the other experimental conditions made this mistake.

The other four critical questions failed to show significant effects for this chi-square analysis, $\chi^2(1, N = 45) < 2$, *ns*. What is important to note, however, is that these same four questions failed to show *any* effect of misleading information, even when the expertise manipulation was disregarded. Chi-square tests were performed on the data for these four questions collapsed across the expertise conditions. This analysis is consistent with the comparisons made in previous experiments of the misleading question effect; the frequency of error among subjects who were asked misleading questions was compared to the frequency of error among subjects who were asked unbiased questions. None of the four questions analyzed in this way showed significant differences in error for misleading versus unbiased questions, $\chi^2(1, N = 45) < 2$, *ns*. These results are consistent with the findings of Dritsas and Hamilton (1977) and Yuille (1980) that were reported earlier: Not all questions are subject to the deleterious effects of misleading information on subsequent accuracy.

Discussion

On the critical question for which misleading information effectively impaired subsequent accuracy, effect was limited, as predicted, to subjects who were given the misleading information by a knowledgeable questioner. The same misleading information presented by a naive questioner did not decrease subjects' subsequent accuracy, with the error rate in this condition being equivalent to the error rate obtained in the two unbiased question conditions.

Experiment 2 was undertaken to address two issues raised by our Experiment 1 results. First, only one of the critical questions tested in Experiment 1 exhibited an error rate high enough to test our hypotheses regarding the role of questioner expertise in obtaining the misleading question effect. Although our results for this question were very strongly in the predicted direction ($p < .01$), we felt it was important to establish the

generality of the findings by replicating the study with a new crime clip and a new set of questions.

Second, there is a seemingly large difference in the percentage of subjects who made errors in the two unbiased question conditions: 43% when asked by a knowledgeable questioner, and only 25% when asked by a naive questioner. As we reported earlier, this difference did not approach significance, $z(\infty) = 0.74$, *ns*, but the number of subjects in each of these cells was small. In Experiment 2 we increased the number of subjects in each of these two conditions to allow a more adequate test of the possibility of a main effect of expertise.

Experiment 2

Method

Subjects

A total of 60 undergraduates enrolled in introductory psychology participated individually in Experiment 2; 15 subjects were randomly assigned to each of the four conditions of the experiment. Data from 2 subjects were excluded from the experiment because they suspected that the confederate was not really another subject. Two additional subjects, randomly assigned, replaced those who were excluded.

Procedure

The procedure for Experiment 2 was the same as that used in Experiment 1, with one exception. We discovered that some subjects in the naive questioner conditions failed to absorb the information that the questioner had no knowledge of the crime. Thus, to make the questioner's naivete in these conditions as salient as his knowledgeability in the other conditions, the following interchange took place:

Experimenter (to confederate): To make your job (of questioning) a little easier, I will tell you that the crime (subject) will witness is a bank robbery in which a hostage is taken.

Confederate (interrupting experimenter's next statement): Wait, you mean I don't get to see the clip?

Experimenter: No, you'll be in the other room writing your questions while (subject) watches the clip. All you get to know about it is that it's a bank robbery in which a hostage is taken.

The addition of the confederate's question and the experimenter's response was quite effective in making the questioner's naivete salient to subjects.

Materials

The videotaped crime clip used in Experiment 2 was an excerpt from a police training film that lasted approximately 4 min. This clip also depicted a bank robbery in which a hostage was taken, but it was not the same clip as that used in Experiment 1.

The critical questions in the interrogation stage of Experiment 2 were as follows (with the unbiased versions of the questions in parentheses):

What did the fourth robber look like? (Was there a fourth robber?)
Was the teller who was taken hostage a man or a woman? (Was the person who was taken hostage a man or a woman?)
How many shots did he fire? (Did he fire any shots?)
Which robber was carrying the bag with the money in it? (Were any of the robbers carrying a bag with money in it?)

As in Experiment 1, one half the subjects were asked the misleading form of the critical questions and the other half were asked the unbiased form; likewise, one half the subjects were questioned by a knowledge-

Table 3
Patterns of Error in Experiment 2

Questioner type	Type of Question	
	Misleading	Unbiased
Percentage of subjects making errors		
Knowledgeable	73	40
Naive	53	40
Mean percentage error per subject		
Knowledgeable	41	13
Naive	18	13

able questioner and the other half by a naive questioner. Following a 20-min filler task unrelated to the experiment, all of the subjects completed a questionnaire measuring their accuracy for the critical details. Questions of interest on this questionnaire were

How many robbers were there? Answer the following about the robber in the red checked jacket who first noticed the police officer: [filler questions] Did he fire any shots? The hostage was a (check one) teller, other bank employee, customer, police officer. Were any of the robbers carrying a moneybag?

Results

As in Experiment 1, contrasts were performed on the arcsine transformed proportions of subjects making errors in each of the four conditions (see Table 3). Consistent with the results of Experiment 1, (a) a significantly higher proportion (73%) of subjects made errors when asked misleading questions by a knowledgeable questioner than when questioned in any other condition (40%–53%), $F(1, \infty) = 3.93$, $p < .05$, and (b) the proportion of subjects making errors when asked misleading questions by a naive questioner did not differ significantly from the proportions obtained for the unbiased question groups, $F(1, \infty) = 0.68$, *ns*. Thus, subjects who were asked misleading questions by a knowledgeable questioner were significantly more likely to make errors in their subsequent reports of the crime than subjects who were asked unbiased questions. Subjects who were asked misleading questions by a naive questioner, on the other hand, did not exhibit such decreases in subsequent accuracy, relative to subjects asked unbiased questions.

The third orthogonal contrast tested for differences in the proportions of subjects making errors in the two unbiased question groups. These two groups yielded identical error rates, so the contrast was nonsignificant, $F(1, \infty) = 0.00$, *ns*. Thus, it appears that the discrepancy obtained in these two cells in Experiment 1 was not in fact due to a real effect of questioner expertise. As explained earlier, one of our objectives in conducting Experiment 2 was to test for a possible main effect of questioner expertise. Thus, an additional contrast for this main effect was tested, but was not significant, $F(1, \infty) = 0.66$, *ns*. These results indicate that an expert who asks unbiased questions does not affect witnesses' error rates.

Additional analyses on the Experiment 2 data examined differences in the average error rate per subject in each of the four conditions of the experiment. Table 3 shows the mean per-

centage of critical questions answered incorrectly in each cell. Consistent with the results reported earlier, the mean proportion of critical questions answered incorrectly for subjects who were asked misleading questions by a knowledgeable questioner is significantly higher than the means for the other three conditions, $F(1, \infty) = 11.35, p < .002$. Furthermore, the error rate obtained for subjects who were asked misleading questions by a naive questioner did not differ significantly from the error rates for the two unbiased question groups, $F(1, \infty) = 0.52, ns$, which in turn did not differ significantly from each other, $F(1, \infty) = 0.00, ns$.

Again, we performed separate chi-square analyses on the data for each critical question. Because our contrast analysis revealed no significant differences in error rates for subjects who were asked misleading questions by a naive questioner and subjects who were asked unbiased questions, these conditions were combined for the chi-square analyses. Thus, our chi-square tests compared the frequency of error for subjects who were asked misleading questions by a knowledgeable questioner to the frequency of error in the other three conditions combined.

Significant chi-square values were obtained for two of the critical questions: "Was the teller who was taken hostage a man or a woman?" yielded $\chi^2(1, N = 60) = 10.72, p < .01$, and "Which robber was carrying the bag with the money in it?" yielded $\chi^2(1, N = 60) = 17.31, p < .001$. Thus, for these two questions, the frequency of error for subjects who were asked misleading questions by a knowledgeable questioner significantly exceeded the frequency of error for the other three conditions.

The remaining two questions failed to produce significant chi-square values, $\chi^2s(1, N = 60) < 1, ns$. We then tested these questions to see if the misleading information significantly affected subjects' accuracy when questioner expertise was disregarded. Comparing the error rate of subjects who were asked misleading questions to the error rate of subjects who were asked unbiased questions (collapsed across expertise conditions), we found no significant effects of misleading information for these two questions, $\chi^2s(1, N = 60) < 1, ns$.

As in Experiment 1 then, misleading information did not always interfere with accurate recall. However, for those questions which *do* show deficits in accuracy due to misleading information, we consistently find that the witness's assumption of questioner expertise is critical. The witness's perception that the questioner already knows about the crime determines whether the misleading information will impair subsequent accuracy.

Discussion

Two important findings emerged from the experiments reported here. The first is the critical moderating role played by questioner expertise in determining the effects of misleading information on accuracy. Our results indicate that the power of a misleading question to distort a listener's memory is not simply a matter of semantics or sentence construction, but involves the listener's perception of the social context.

As in previous research, misleading questions generated incorrect answers when the witness could assume that the interrogator already knew a great deal about the crime. In a situation like this, unless the witness has a very clear memory that contradicts the facts presupposed by the knowledgeable questioner, he

or she is not likely to doubt the accuracy of those facts, and will later remember them as part of the event. But when the witness knows that the interrogator is ignorant of the facts, misleading questions have no effect. In a situation like this, the witness is the primary authority on the crime, and is unlikely to accept false presuppositions in the first place or to recall them later. Subjects who were asked misleading questions by an interrogator they knew to be naive were as accurate as subjects who were not asked misleading questions at all.

These findings do not raise questions about the validity of the misleading question effect so amply documented by Loftus and others, only about its generality. That facts presupposed in a question influence the respondent is not simply a cognitive phenomenon; it is also a social phenomenon. Social psychologists have known for decades that a direct communication, intended to persuade, will be more effective in changing attitudes if the source is seen as credible (Hovland & Weiss, 1951). A situation in which one person questions another is at least as much a social interaction as a situation in which one person attempts to persuade another, and it is not surprising that characteristics of the communicator should be influential in both. The same communicator qualities that affect the power of direct communications affect the power of indirect ones.

The second important finding highlighted by these experiments is one that has been suggested before but may be more general than is typically assumed: Memories for some facts are relatively immune to alteration by the presentation of misleading information. As we described earlier in this article, it appears that only memories that are somewhat indefinite are subject to distortion on the basis of subsequent information (Dritsas & Hamilton [cited in Loftus, 1979]; Yuille, 1980). The differential effectiveness in distorting recall of the critical questions used in the two experiments reported in this article lends further support to this uncertainty hypothesis. Also consistent are the findings of a study reported by Marquis, Marshall, and Oskamp (1972) that the accuracy of eyewitness testimony varies as a function of the difficulty of the questioned item, with higher error rates obtained for "difficult" items than for "easy" ones.

Although the potential exists to impair a witness's memory for a crime by asking misleading questions, this type of memory distortion is subject to certain preconditions. Two of these preconditions are the expertise of the person presenting the misleading information and some uncertainty about the facts. The mere presentation of misleading information is not sufficient to decrease a witness's accuracy.

Applications

The findings of the experiments reported here also have practical implications for police questioning. Given that accuracy is the goal in questioning eyewitnesses, our results suggest that police officers should avoid suggesting to witnesses that they already know something about the crime, because under these circumstances, if a misleading question is inadvertently asked, the accuracy of the witness's report is likely to suffer. If the police are successful in convincing witnesses that they are ignorant of what took place, misleading questions will generally not result in decreased witness accuracy.

But how do witnesses generally see the police officers who

question them? One possibility is that the police are perceived as experts, using the witness to confirm and elaborate a scenario they already have in mind. This may occur because the police present themselves as more knowledgeable than they really are—for example, because they have a suspect in mind but no corroboration, because they have already obtained information about the crime from other witnesses, or because of their general knowledge of similar crimes. On the other hand, it may occur simply because the witness believes in the expertise of the police, regardless of the behavior of the particular officer who is asking the questions. In either case, the phenomenon described by Loftus would represent a real threat to the accuracy of the witness's report, and could increase the number of false arrests and convictions. Another possibility, however, is that witnesses generally believe that the police are asking questions because they really don't know what happened. When this is the case, it is much less likely that an officer's assumption about what "must have happened" could distort the witness's memory of what did happen.

In any case, when the police are seeking information they should go out of their way to convince the witness that they know little or nothing of what happened. But in order to know how serious a problem the misleading question effect currently is, it is important to move from the lab to the police station. We need to find out, first and most obviously, what kinds of questions police actually ask. How common are misleading questions, incorporating presuppositions based on specific information about the crime or general knowledge of similar crimes? Second, we need to know whether witnesses see the police as *knowing* or as *seeking* the truth. Finally, we need to know what it is the police do that may create a false impression of expertise in some or most witnesses, and what they can do to counteract it.

References

- Clifford, B. R., & Scott, J. (1978). Individual and situational factors in eyewitness testimony. *Journal of Applied Psychology*, 63, 352-359.
- Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly*, 15, 635-650.
- Langer, E. J., & Abelson, R. P. (1972). The semantics of asking a favor: How to succeed in getting help without really dying. *Journal of Personality and Social Psychology*, 24, 26-32.
- Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology*, 7, 560-572.
- Loftus, E. F. (1977). Shifting human color memory. *Memory and Cognition*, 5, 696-699.
- Loftus, E. F. (1979). *Eyewitness testimony*. Cambridge, MA: Harvard University Press.
- Loftus, E. F., Altman, D., & Geballe, R. (1975). Effects of questioning upon a witness's later recollections. *Journal of Police Science and Administration*, 3, 162-165.
- Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory*, 4, 19-31.
- Loftus, E. F., & Palmer, J. P. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, 13, 585-589.
- Loftus, E. F., & Zanni, G. (1975). Eyewitness testimony: The influence of the wording of a question. *Bulletin of the Psychonomic Society*, 5, 86-88.
- Marquis, K. H., Marshall, J., & Oskamp, S. (1972). Testimony validity as a function of question form, atmosphere, and item difficulty. *Journal of Applied Social Psychology*, 2, 167-186.
- McGuire, W. J. (1969). The nature of attitudes and attitude change. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (pp. 136-314). Reading, MA: Addison-Wesley.
- Mosteller, F., & Tukey, J. W. (1949). The uses and usefulness of binomial probability paper. *Journal of the American Statistical Association*, 44, 174-212.
- Yuille, J. C. (1980). A critical examination of the psychological and practical implications of eyewitness research. *Law and Human Behavior*, 4, 335-345.

Received March 6, 1986

Revision received September 13, 1986 ■

Perceptual and Preferential Discrimination Abilities in Taste Tests

Moshe M. Givon

Recanati Graduate School of Management
Tel-Aviv University, Tel-Aviv, Israel

Arieh Goldman

Jerusalem School of Business Administration
Hebrew University, Jerusalem, Israel

We investigated the relation between perceptual and preferential discriminative ability in taste tests, using a convergence of methods to construct measures of ability. A sample of 175 respondents participated in a taste test of tomatoes. The sample included graduate students, university staff members, high school teachers, and army trainees. Results indicate (a) that perceptual and preferential discriminative abilities are related but they do not precondition each other; (b) that only 16% of the respondents were good discriminators in both perception and preference, 24% were perceptual discriminators, and 42% were preferential discriminators; and (c) that discriminators and nondiscriminators exhibit different preferences.

Product development in the food industry often involves comparative testing. Consumers are offered two or more unlabeled products and are asked to compare them on some basis such as taste, smell, or appearance. The comparative testing involves two types of tasks—perceptual discrimination and preferential testing (Batsell & Wind, 1979).

Perceptual discrimination focuses on whether a product is actually perceived as being different from some existing or reference item. This is a pertinent question when a new brand is created or an existing one changed. A firm developing a sweeter soft drink wants to know if consumers do indeed find it sweeter. Or if a company substitutes an expensive ingredient for a cheaper one, it wants to know if consumers will notice a difference in taste. If consumers cannot discriminate between the new product and its competitors or cannot see that a change has been made, the product is unlikely to reach the market. The most common method used to measure perceptual discrimination ability is the triangle test (Moskowitz, 1983, 1985). Respondents evaluate three items, two of which are identical, and are asked to indicate the odd one. Those who are successful in repeated trials are considered to be discriminators.

In preference testing, consumers are presented with two or more products and are asked to choose the one they prefer. The tasks used often involve paired comparison or direct ordering. Preferences are viewed as decision-making dispositions and used as an indicator of the future acceptance of the new product.

The double triangle test (a triangle test repeated twice) coupled with a single paired comparison is often used to screen out

nondiscriminators in preference tests (Buchanan & Morrison, 1985; Day, 1974; Moskowitz, Jacobs, & Firtle, 1980). The use of the triangle test as a measure of discrimination in preference tests is based on Irwin's (1958) view that respondents cannot have a true preference if they cannot discriminate.

The main argument against using a measure of perceptual discrimination as an indirect test of discrimination in preference tests is based on Zajonc's (1980) views. Traditional psychological models, which Irwin followed, imply that preferences are formed and exhibited after considerable prior cognitive activity. Zajonc raised the question of whether objects must be cognized before they can be evaluated. It is possible that we can like something before we know much about it. Furthermore, Zajonc suggested that the stimulus features used in perceptual tasks may not be useful in the evaluation tasks. He believes that a view of two independent systems is more plausible than the one that relegates affect to a secondary role, mediated and dominated by cognition. In the present context Zajonc's argument implies that perception is not a precondition for preference, and therefore, measures of one should not be used as indirect measures of the other. This hypothesis is tested later.

Repeated paired comparisons is the second major approach used to evaluate discrimination ability in preference tests (Buchanan & Morrison, 1985; Greenberg & Collins, 1966; Hyett & McKenzie, 1976; Moskowitz et al., 1980). Respondents are given two products and are asked to select the one they prefer. Those who choose the same product in repeated tests are considered to be preferential discriminators. The crucial assumption is that a subject's preferences are consistent for all trials in the test (Buchanan & Morrison, 1984; Day, 1974). But boredom, variety seeking, and learning may lead subjects to change their choices, even if they discriminate (McAlister & Pessemier, 1982).

Another consideration is that the production process for many foods is not fully standardized, resulting in high product variability, particularly in the case of fresh, canned, and frozen fruits and vegetables. The problem may also appear among processed foods (e.g., soft drinks, cereals, jams, cheese, beer), wherein items from different production runs can vary in taste.

Preparation of this article was supported in part by the Israeli Ministry of Agriculture and the Israel Institute of Business Research, Tel-Aviv University.

We gratefully acknowledge the assistance of two anonymous reviewers for their comments on a draft of this article.

Correspondence concerning this article should be addressed to Moshe Givon, Recanati School of Management, Tel-Aviv University, Ramat-Aviv, Tel-Aviv 69978, Israel.

In these cases it is difficult to ensure that repeated tests will be conducted on the same item.

Assumptions about consistency of preferences and stability of discrimination ability over trials can greatly limit the usefulness of the repeated paired-comparisons design. Another problem is a more practical one. Repeated tests do not yield additional information about the nature of preferences; they are only needed for testing discrimination. A method that will simultaneously yield insights into the nature of preferences and also serve as a measure of discrimination is superior.

In this article we deal with two issues that are central to testing discrimination. The first is the need to distinguish clearly between discrimination in perceptions and in preferences and their measurement. The second issue is the following: Current procedures for testing discrimination are based largely on the idea of the consistency of responses in repeated tests. In the area of preferences, this is the only criterion for discrimination. There is a need for testing procedures that do not rely on consistency of preferences.

The approach proposed here for testing discrimination is based on convergence of taste evaluations generated by different methods. Convergence of alternative methods has been discussed extensively in the psychometric literature on validity. We suggest that it can also be used to measure discrimination. Researchers and practitioners use a number of alternative methods for measuring perceptions and preferences. In the case of preferences, for example, rank ordering, paired comparisons, and distance from ideal point are used to measure the preference construct. Although each method aims at a different aspect of the construct, in practice they are considered interchangeable. Agreement between the choices made or implied when two alternative methods are used can be a measure of discrimination. It is useful in practical applications because preferences and perceptions generated from consumers whose judgments vary with the method used (nondiscriminators) should be evaluated differently from those of consumers who consistently offer the same judgments.

We view the measurement of discrimination as an evaluation of the validity and reliability of the data generated from taste tests. The consistency criterion, dominating current measurement practices in testing preferences, represents only one aspect of reliability. The convergence criterion, suggested here, captures another dimension. Researchers should select the appropriate approach, given the specific test circumstances and purposes.

In the following sections we describe an experimental taste test, develop measures of perceptual and preferential discrimination, and investigate the relation of the measures.

Method

Subjects and Materials

The study, conducted in Israel, involved 175 respondents who regularly ate tomatoes (about 75% ate tomatoes daily; average consumption was 1.2 tomatoes per day). As the objective of the study was methodological, convenience samples were drawn from a number of population groups: graduate students and staff from a major university, teachers at a large high school, and trainees in two advanced army courses. Respon-

dents evaluated five tomatoes (two of the same variety and three of different varieties). All of the tomatoes were recently developed, were grown in the winter in hothouses, and were not commercially available at the time of the study.

Procedure

On his or her arrival in the test room, the respondent was shown to a seat facing the interviewer at a table. Five plates were placed on the table, each containing a large number of small pieces cut from one medium-sized tomato. Each plate was identified with a two-digit code, seen only from the interviewer's side. A pitcher of mineral water and a plate of crackers (low sodium) were also on the table, and respondents were encouraged to rinse their mouths, eat a cracker, and rest "whenever you feel you're losing your ability to taste." Most respondents completed the tasting tasks using only about 60% of the tomato pieces on each of the plates (a total of about three medium-sized tomatoes). Tests were generally conducted between breakfast and lunch and lasted an average of 19 min. Based on comments made during the test and in the debriefing session, respondents found the taste tasks easy to perform and did not suffer from sensory fatigue or boredom. Although some did complain about the relatively large amounts of tomatoes they had to eat, only 5 participants left in the middle of the experiment. All of the respondents were told that they were tasting five new varieties of tomatoes, especially developed for export to Europe and the United States, but not yet available commercially. Supposedly, the purpose of the test was to help Israeli exports by selecting the best tasting tomato for commercialization.

Tasks

The test tasks were designed to measure the perceptions and preferences of each respondent. Two methods were used for each—similarity judgment and attribute rating for measuring perceptions, and paired comparison and rank ordering for preferences. For validation purposes, information was collected about ideal points as well as socioeconomic and consumption data to check for possible segmentation.

The sequence of experimental tasks was as follows.

Similarity judgment. All 10 possible combinations of three tomatoes were tasted sequentially by the respondent. For each triad, the respondent indicated which two were the most similar and which two were the most dissimilar. Note that if two of the tomatoes in a triad are identical, the similarity judgment becomes a triangle test. Thus, the triangle test is a special case of similarity judgment in triad. We believe it is appropriate to use similarity judgment in situations in which absolute control over stimuli variance is not assured, as in the case with the agricultural products studied here.

Paired comparison. The five study tomatoes yielded 10 possible combinations of pairs. For each pair, respondents picked their preferred tomato.

Preference ranking. Respondents rank ordered the tomatoes according to preference.

Consumption and buying habits. Respondents answered a set of questions about their tomato consumption and buying habits.

Attribute rating. Respondents evaluated all of the tomatoes on a 5-point semantic differential scale (e.g., *very sweet* to *not sweet at all*), on each of 12 attributes (see Figure 1). The attributes were generated from the literature and from in-depth interviews previously conducted.

Ideal point. Respondents indicated on a 5-point scale the desired degree of each attribute in their ideal tomato and then showed how important it was for them that the ideal tomato contain this degree of the attribute.

Socioeconomic information. Respondents answered questions on their socioeconomic status and were debriefed.

Measures of Discrimination

Two measures of discrimination were developed, one for perceptual and one for preferential discrimination. Both are based on the principle of convergence of alternative methods.

Perceptual discrimination ability. Similarity judgments in triads and attribute ratings were used to measure respondents' relative similarity in perceptions of the tomatoes. The discrimination measure we developed is based on the assumption that a pair of tomatoes that is judged more alike than another pair should also be closer in attribute profiles. The dissimilarity of attribute profiles of any pair was calculated as the sum of absolute differences on the 12 attributes, that is, city-block distance. Each triad produced three paired comparisons (ij vs. ik, ik vs. jk, and ij vs. jk); altogether there were 30 such comparisons in the 10 triads. The proportion (D_1) of paired comparisons with the same relation, according to both similarity judgment and attributes evaluations, was calculated for each respondent.

Preferential discrimination ability. The five test tomatoes formed 10 pairs. Preferences for each pair were obtained by two methods: the paired comparison test and the preference rank ordering task. The proportion of identical paired relations in both methods was calculated for each respondent and was labeled D_2 .

Results

Perceptual Discrimination

Respondents' scores of perceptual discrimination ranged from .2 to 1, with a mean of .61.

A major problem with the convergence-based measure of discrimination is that there is always the possibility that the two methods used to measure the construct (e.g., perception) have a poor convergent validity. In this case, the responses of dis-

criminator should not be correlated; a high correlation might indicate guessing. To see whether D_1 does measure perceptual discrimination, we check to see if the discriminators show a higher ability to detect the two tomatoes of the same variety. If the perceptual discriminators do a better job than the nondiscriminators, our confidence in D_1 as a measure of perceptual discrimination increases.

The respondents were placed in three groups according to their D_1 score: $.0 \leq D_1 \leq .5$; $.5 < D_1 \leq .7$; $.7 < D_1 \leq 1$. The first group were considered to be nondiscriminators, whereas those with a score of $D_1 > 0.7$ were regarded, for the purpose of the analysis here, as discriminators.

The five test tomatoes were coded as 26, 42, 53, 61, and 84. Tomatoes 42 and 84 were of the same variety, whereas the others were each different. Because every respondent tasted five different tomatoes and because tomatoes of the same variety or even of the same plant can vary widely, no intraindividual analysis is possible. On the average, however, tomatoes of the same variety should be more alike than those of different varieties.

Comparing the differences of average profiles for tomatoes of the same variety (numbers 42 and 84) with those of tomatoes of different varieties, we expected that the latter would be larger for discriminators and about the same for the nondiscriminators. The absolute differences of average profiles are given in Figure 1 for discriminators ($D_1 > .7$) and in Figure 2 for nondiscriminators ($D_1 \leq .5$). With Tomato 42 chosen as the basis of comparison, the absolute difference between average profiles of numbers 42–84 gives the "natural-standard" difference between tomatoes of the same variety. The results in Figures 1

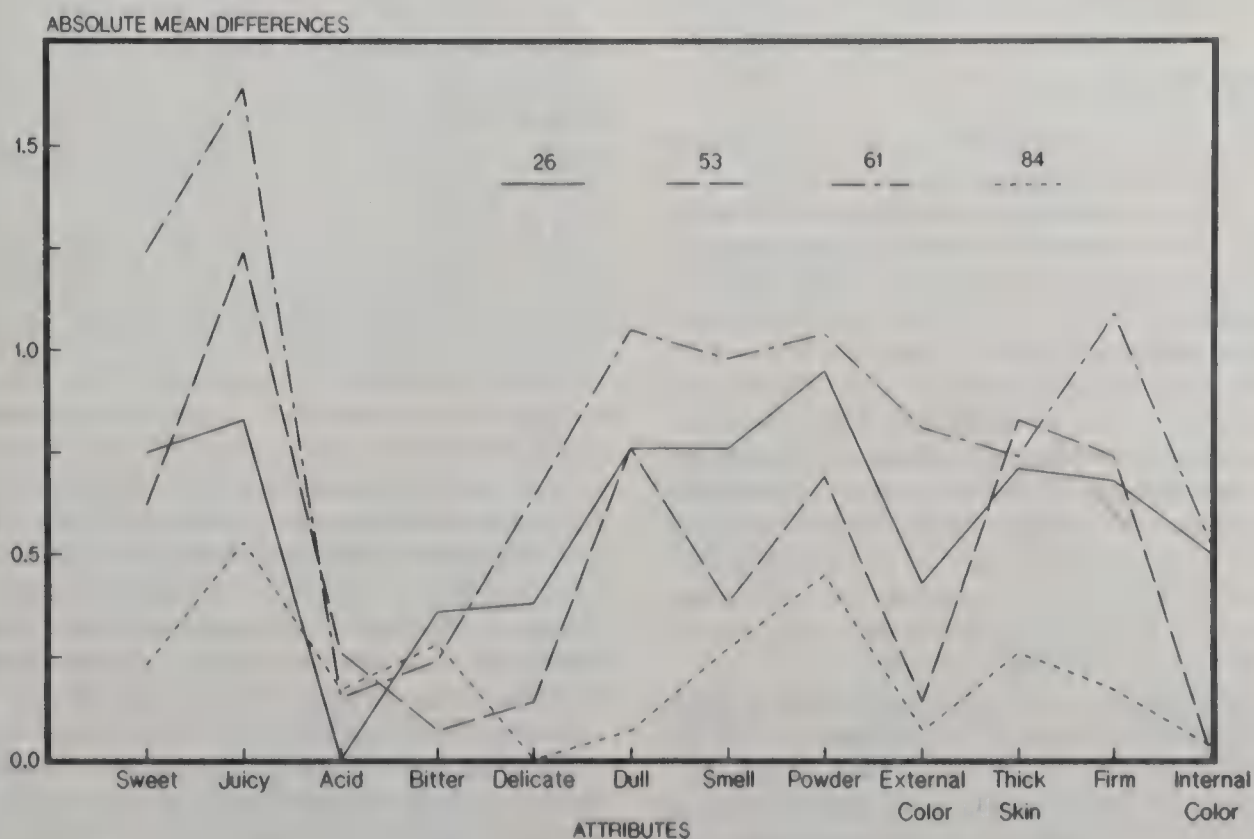


Figure 1. Discriminators' attributes profiles of absolute differences from Tomato 42.

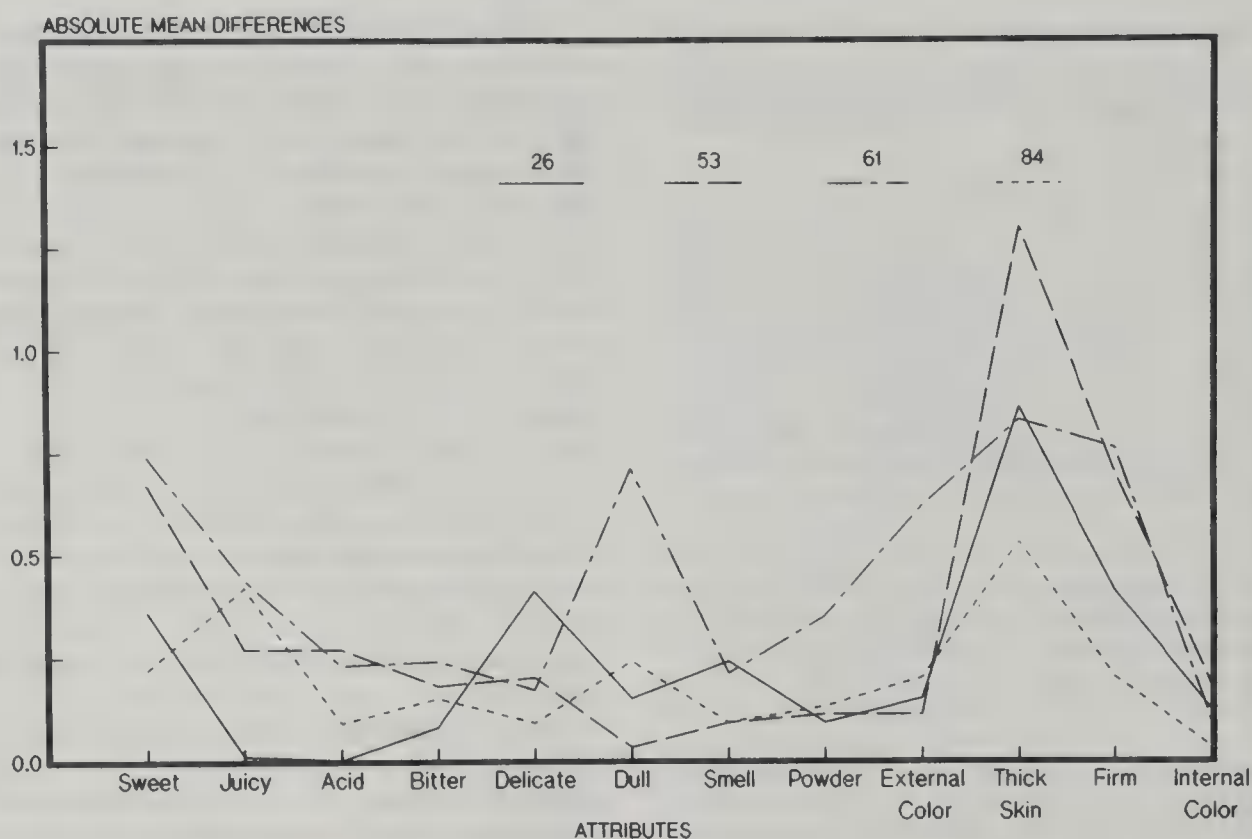


Figure 2. Nondiscriminators' attributes profiles of absolute differences from Tomato 42.

and 2 show that the perceptual discriminators did a good job in distinguishing among the different varieties in all attributes except acidity and bitterness, whereas the nondiscriminators did a poorer job. We concluded that D_1 does indeed measure perceptual discrimination.

Preferential Discrimination

Respondents' scores of preferential discrimination ranged from 0 to 1, with a mean of .67 and a median at .70.

The validity problem discussed in the context of D_1 with a convergence-based measure of discrimination applies here, too. There is, therefore, a need to validate that D_2 does measure preferential discrimination. Here again, we could use the fact that two tomatoes (numbers 42 and 84) were of the same variety and therefore should be equally preferred. Unfortunately, both discriminators and nondiscriminators are expected to give equal average preference ranks to both tomatoes. Nondiscriminators should do so because they rank randomly, whereas discriminators can taste the similarity. Indeed, the differences in average ranks of the two tomatoes were very small and statistically insignificant for both groups. We are left then with the face validity of preference ranking and paired comparisons, relying on their extensive use to measure preferences.

Measuring preferential discrimination has no practical value if the choices of discriminators and nondiscriminators are the same. This indeed was the case in many earlier studies (e.g., Greenhalgh, 1976; Moskowitz et al., 1980; Penny, Hunt, & Twyman, 1972). If a preferential discrimination measure is to be of practical use, it should be related to actual choice in the

following way: Nondiscriminators can be expected to distribute their choices more evenly among the alternatives, whereas discriminators should choose according to their real preferences that may or may not be evenly distributed.

The information about consumers' ideal tomatoes provides insight into the preference structures of discriminators ($D_2 > .7$) and nondiscriminators ($D_2 \leq .5$). The mean of each group on each attribute of their ideal tomato was calculated. The 12 separate t tests for equality of means showed that the two groups differed significantly ($p = .01$) only with respect to their ideal level of sweetness. However, the ideal tomato of both groups was sweeter than were any of the five test tomatoes. Given the similarity of their ideal tomatoes, we checked for differences in choices by comparing first choices of discriminators to those of nondiscriminators. The results are shown in Table 1. A chi-square test for independence showed a significant relation between preferential discrimination by respondents and their choice of preferred tomatoes, $\chi^2(8, N = 175) = 18.67$, ($p < .02$). We compared each group's actual distribution of first choices to the hypothetical uniform distribution (where 20% of the choices go to each of the five tomatoes) by a chi-square test. The results indicate that the null hypothesis of uniform choice distribution could not be rejected for the nondiscriminators, whereas it should be rejected for discriminators ($p = .01$).

We discussed earlier our objections to the use of a perceptual discrimination measure for evaluating preferential discrimination. To assess the relation between these two measures, we correlated D_1 and D_2 . Their correlation ($r = .34$) was significant ($p = .01$), but not high.

Table 1
*Number of Respondents With Preferential
 Discrimination by First Choice*

First choice	Preferential discrimination (D_2)		
	Low (.00-.50)	Medium (.51-.70)	High (.71-1)
Tomato 26	15	18	9
Tomato 42	9	7	22
Tomato 53	10	4	15
Tomato 61	7	9	6
Tomato 84	10	12	22

Note. $\chi^2(8, N = 175) = 18.67, p < .02$.

We checked to see whether consumption habits (e.g., main usage form, frequency and quantity used, shopping location, price paid) or demographic variables could be used to segment respondents by their discrimination ability. We either correlated or cross tabulated these variables with the D_1 and D_2 scores. Although the respondents came from different population groups, there was no statistically significant relation between discrimination ability and any of these variables.

Discussion

The present data indicate that a perceptual discrimination measure should not be used to evaluate preferential discrimination. Although both measures are related, neither preconditions the other. This result casts doubt on the widely accepted assumption that perceptual discrimination is a necessary precondition for preference discrimination. The prevalent practice of using the triangle test to evaluate preferential discrimination should be discontinued. Instead, there is a need to develop appropriate procedures for directly measuring preferential discrimination. This issue should now receive the attention it deserves in the taste-testing literature.

Our study found that many respondents had poor discriminative ability. Very few (16%) were good discriminators in both perception and preference; only 24% were perceptual discriminators; and some 42% were preferential discriminators. These findings are similar to those reported in earlier studies (Day, 1974; Greenberg & Collins, 1966; Gruber & Lindberg, 1966; Moskowitz et al., 1980). Although comparisons are difficult because of differences in products, populations, and the methods used for evaluating discrimination, the general pattern seems quite clear—a significant number of consumers are nondiscriminators in taste. Given the magnitude of this problem, the current practice of largely ignoring the discrimination issue in studies of consumers' tastes (Moskowitz, 1983, 1985) cannot be justified. In every taste test, practitioners should directly evaluate discrimination ability of their respondents.

One problem that makes it difficult to incorporate a discrimination evaluation as an inherent part of each taste test is the large cost in respondent time and effort involved in repeated tests, not to mention the possibility of boredom. The convergence measure suggested here offers an alternative to consistency-based discrimination measures. Besides its usefulness in

case of nonstandardized products, it is easy to use, does not suffer from the various shortcomings of the repeated tests, and does not require special purpose testing as repeat measures. Because preference studies often involve two methods for measuring preferences, preferential discrimination measurement can become an integral part of the study. Note that although the research design in this study required respondents to perform a large number of taste tasks, many of the tasks were required for the comparison of perceptual and preferential discrimination abilities. In a normal testing situation, the practitioner would not need to do this comparison, greatly reducing the number of test tasks.

Most of the earlier discrimination studies reported that the choices made by the discriminators and nondiscriminators were similar (Day, 1974; Moskowitz et al., 1980). If so, this weakens the argument for conducting discrimination tests. Indeed, these results may have legitimized the tendency, mentioned before, to omit discrimination testing altogether. In this context, the findings reported here, that preferential nondiscriminators do make different choices from discriminators, is of major importance. Underscoring the need to assess respondents' discrimination abilities in every specific taste test, it supports Buchanan and Morrison's (1984) contention that the results previously reported in the literature reflect spurious classification and that when discrimination is properly measured, discriminators and nondiscriminators do perform differently.

Our study does suffer from obvious limitations. The experiment involved only one product and a relatively small sample, limiting its generalizability. Replication with different products and other samples is needed. To be able to generalize about the relation between perceptual and preferential discriminative ability, a similar analysis should be conducted using consistency-based measures of discriminative ability (e.g., repeated triangle tests and repeated paired comparisons).

A number of issues also deserve additional research. What is the relation between convergence-based measures of discrimination ability and measures based on the consistency criterion? An analysis of this relation will provide insights into discrimination and help validate its measures. We could not identify use of demographic correlates of discrimination, which raises the question: To what extent is discriminative ability a stable characteristic of the individual rather than a result of situational factors such as mood, health, fatigue, and foods previously eaten? Will those who display a high discriminative ability with one product also discriminate well in other product classes? The answers have major implications for current taste-testing practices that seem to assume that discrimination ability is a characteristic of the individual that can be generalized to any food product.

References

- Batsell, R. R., & Wind, Y. (1979). Product testing: Current methods and need developments. *Journal of the Market Research Society*, 22, 115-137.
- Buchanan, B. S., & Morrison, D. G. (1984). Taste tests, psychophysical issues in comparative test design. *Psychology & Marketing*, 1, 69-91.
- Buchanan, B. S., & Morrison, D. G. (1985). Measuring simple prefer-

- ences: An approach to blind, forced choice product testing. *Marketing Science*, 4, 93-109.
- Day, R. L. (1974). Measuring preferences. In R. Ferber (Ed.), *Handbook of marketing research* (Part 3, pp. 101-125). New York: McGraw-Hill.
- Greenberg, A., & Collins, S. (1966). Paired comparison taste tests: Some food for thought. *Journal of Marketing Research*, 3, 76-80.
- Greenhalgh, C. (1976). Discrimination tests and repeated paired comparison tests. *Journal of the Market Research Society*, 18, 214-215.
- Gruber, A., & Lindberg, B. (1966). Sensitivity, reliability, and consumer taste testing. *Journal of Marketing Research*, 3, 235-238.
- Hyett, G. P., & McKenzie, J. R. (1976). Discrimination tests and repeat paired comparisons tests. *Journal of the Market Research Society*, 18, 24-31.
- Irwin, F. W. (1958). An analysis of the concepts of discrimination and preference. *American Journal of Psychology*, 11, 152-163.
- McAlister, L., & Pessemier, E. (1982). Variety seeking behavior: An interdisciplinary review. *Journal of Consumer Research*, 9, 311-322.
- Moskowitz, H. R. (1983). *Product testing and sensory evaluation of foods*. Westport, CT: Food & Nutrition Press.
- Moskowitz, H. R. (1985). *New directions for product testing and sensory analysis of foods*. Westport, CT: Food & Nutrition Press.
- Moskowitz, H. R., Jacobs, B., & Firtle, N. (1980). Discrimination testing and product decisions. *Journal of Marketing Research*, 17, 84-90.
- Penny, J. C., Hunt, I. M., & Twyman, W. A. (1972). Product testing methodology in relation to marketing problems—A review. *Journal of the Market Research Society*, 14, 1-29.
- Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, 35, 151-175.

Received August 1, 1986 ■

Role of Efficacy Expectations in Predicting the Decision to Use Advanced Technologies: The Case of Computers

Thomas Hill and Nancy D. Smith
University of Tulsa

Millard F. Mann
University of Kansas

The complexity of innovations has long been recognized as a factor affecting the rate of adoption. We investigated the relation between sense of efficacy regarding computers and people's readiness to use them. Using structural equation modeling procedures (LISREL) in Study 1, we showed the hypothesized relation between efficacy beliefs with respect to computers and the likelihood of using computers (as measured by subsequent enrollment in computer-related courses) in two independent samples. We demonstrated that beliefs of efficacy regarding computers exert an influence on the decision to use computers that is independent of people's beliefs about the instrumental value of doing so. In Study 2 we extended this finding by showing that, consistent with Bandura's research on the personal efficacy construct, previous experience with computers is related to beliefs of efficacy with respect to computers, but that it does not exert a direct independent influence on the decision to use computers. Furthermore, a significant relation was found in Study 2 between general beliefs of personal efficacy and use of other electronic devices. These studies demonstrate the importance of efficacy beliefs in the decision to adopt an innovation.

People react strongly to computers. Many seem convinced that an electronic paradise, wherein all of the work is done by sophisticated electronic gadgets, is just around the corner. Some are compelled by the challenge to be part of this new age, to find new computer algorithms, develop fancy graphics, or write more sophisticated programs. For these people the computer's ability to process large amounts of information at high speeds makes it irreplaceable for facilitating a variety of tasks.

However, "techno-phobics" and computer illiterates seem unlikely to attach such value to these machines. They consider computers too complex. They believe that they will never be able to control these devices and prefer to avoid them. One might expect such beliefs to be negatively related to people's intentions to use computers.

Perceived complexity of innovations has long been recognized as a factor inhibiting their diffusion (e.g., LaBay & Kinnear, 1981; Rogers, 1962). In general, investigators have proposed that increased complexity of an innovation requires increased cognitive effort on the part of the adopter, thus decreasing the likelihood of adoption (Dickerson & Gentry, 1983; Hirschman, 1980).

Cognitive laziness may, in fact, adequately explain why some people are reluctant to use computers. However, most teachers who have introduced students to computers will probably agree that novices are often rather frightened by the anticipated inter-

action with the machine, despite their willingness to expend effort.

An alternative explanation is that many people may not believe that they will ever be able to interact successfully with computers, that is, to control them. Research has demonstrated the negative attitudinal and behavioral effects of loss of perceived control (see Seligman, 1975, for a review). Bandura and his associates (see Bandura, 1977; Bandura, Adams, & Beyer, 1977; Bandura & Schunk, 1981) have convincingly demonstrated the role of (lack of) personal efficacy (i.e., the belief that one is able to master a particular behavior) in phobias.

The rapid technological advance from slide rules to calculators to microcomputers may have overwhelmed some people and left them with little sense of efficacy regarding computers. Furthermore, initial experiences are often frustrating and not likely to strengthen the belief that computers can be controlled.

Study 1

The primary purpose of Study 1 was to investigate the relation between people's expectations of being able to control computers (i.e., computer efficacy beliefs) and their decision to use them. We predicted that the more controllable computers are believed to be, the more likely people are to use them. This hypothesis was tested in two samples of male and female college students via linear structural equation modeling procedures (LISREL; Jöreskog & Sörbom, 1978, 1979).

Method

A questionnaire was administered to a sample of 157 female and 147 male undergraduate students enrolled in an introductory psychology class. Male and female respondents were treated as two independent

The authors would like to thank Jack W. Brehm, Jim Fultz, and Lynne Steinberg for their helpful comments regarding this article.

Correspondence concerning this article should be addressed to Thomas Hill, University of Tulsa, Department of Psychology, 600 South College Avenue, Tulsa, Oklahoma 74104.

samples in the analyses, allowing for independent replication of the results. The questionnaire contained items designed to assess efficacy beliefs with respect to computers, items to measure beliefs about the instrumental value of learning about computers, and items designed to measure behavioral intentions to purchase or use computers in the future (i.e., to enroll in computer-related courses in the following semester). In addition, actual enrollment in the following semester's computer courses was assessed 12 weeks after subjects had completed the questionnaire.

Computer efficacy. The items intended to measure computer efficacy beliefs were originally formulated by a group of teachers with experience in introducing students to the time-sharing computer system at a large midwestern state university (Hill & Smith, 1985). It has previously been shown that people who score low on this scale (low sense of efficacy regarding computers) are more easily persuaded by expert communicators to try an innovative computer product than are people who score high on this scale (Hill, Smith, & Mann, 1986). In the Hill et al. study, college students were asked to evaluate an advertisement for a software package that supposedly was designed specifically for students. After watching an advertisement in which "experts" described the product, subjects who scored low on the computer efficacy beliefs scale were more likely to sign up for a trial of this package than were subjects who scored high on this scale. This finding is consistent with the results of the general research concerning persuadability as a function of source expertness and locus of control (e.g., Ritchie & Phares, 1969; Ryckman, Rodda, & Sherman, 1972; see Lefcourt, 1982, or Phares, 1978, for reviews), and reflects favorably on the validity of this four-item scale.

The items used to measure computer efficacy beliefs were as follows: Item 1, I will never understand how to use a computer; Item 2, Only a few experts really understand how computers work; Item 3, It is extremely difficult to learn a computer language; and Item 4, Computer errors are very difficult to fix. Each item was accompanied by a 5-point scale ranging from *totally agree with the statement* (1) to *totally disagree with the statement* (5).

Instrumentality beliefs. The general procedure described by Ajzen and Fishbein (1980) was followed in order to determine the specific benefits that students believe result from learning to use computers: A questionnaire was administered to a sample of 31 male and female college students enrolled in an introductory psychology class. They were asked to write down as many as eight benefits that they believed would result from learning to use computers—in the order of importance for them personally.

Subjects' responses were coded by two independent raters using five categories: job-related benefits (mostly competitiveness in the job market, higher salary, higher status jobs); personal growth (e.g., interest, challenge); entertainment; household management (e.g., financial planning); and other benefits (e.g., being able to teach one's children, being fashionable, facilitating the preparation of manuscripts). All of the respondents reported at least two outcomes that they believed were associated with learning to use computers. The interrater agreement for coding the most important and the second most important outcomes was 90% and 84%, respectively. Subjects' responses strongly suggested that job-related outcomes were the most important benefits they believed would result from learning to use computers. With the exception of 2 students, all of the respondents listed job-related benefits as either most or second most important. It can be concluded from this pilot study that beliefs about job-related benefits from learning to use computers are most salient among students and can be expected to have a strong influence on their decision to learn to use them.

Thus, four items were written for the questionnaire assessing beliefs about the instrumental value of being familiar with computers: Item 5, I will not get as high a starting salary when I graduate if I don't know how to use a computer; Item 6, If I know about computers I can get a higher status job; Item 7, Expertise in computers is of utmost impor-

tance if I want to get a good job; and Item 8, If I don't learn how to use computers it will be difficult to be successful in any professional career. In other research Hill and Smith, 1985, demonstrated that these items measure one, unidimensional construct. Each item was accompanied by a 5-point scale ranging from *strongly disagree* (1) to *strongly agree* (5).

Behavioral intentions. The following three questions assessed respondents' intentions to use computers: Item 9, Are you intending to purchase a personal computer within the next year or so?; Item 10, Are you intending to take computer science classes this semester or during the next semester?; and Item 11, Are you intending to learn a computer language in the near future? All items were accompanied by a 5-point scale: *yes* (5), *probably* (4), *undecided* (3), *probably not* (2), and *no* (1).

Behavior. At the end of the semester (approximately 12 weeks after the administration of the questionnaire), respondents were contacted by telephone. Because it became clear during pilot research that only a few students would purchase a computer within a period of 12 weeks, the decision to enroll in computer science courses (or other courses requiring the use of computers) was assessed as a behavioral indicator of adoption of computer technology. This procedure appears justified because the adoption of computer technology involves learning how to use this technology. At the time of the telephone interview, students had already enrolled for the following semester. Respondents were asked whether they had enrolled in courses requiring the use of computers. The responses were coded dichotomously as *yes* (1) or *no* (0).

Data analysis. The data analysis was performed via structural equation modeling procedures. Specifically, we used LISREL IV (Jöreskog & Sörbom, 1978) and a version of LISREL VI (Jöreskog & Sörbom, 1985) for the IBM/PC (Version VI.9). The microcomputer version is a full implementation of LISREL VI, except that it lacks the option to analyze ordinal scale measures (via polychoric correlation coefficients).

The parameter estimates (and chi-square values) calculated by LISREL are maximum likelihood estimates under the assumption that all variables are normally distributed and not restricted in range. In Study 1, the questionnaires were administered along with several others in a "mass testing" session. We used 5-point scales in order to be consistent with the response format used in other questionnaires. Note that the potential restriction of range may constitute a violation of the assumptions underlying the use of LISREL.

Behavior (i.e., actual enrollment in the following semester's computer courses) was treated as a dichotomous variable. Although the estimation of the maximum likelihood parameters in LISREL assumes interval scale measures, behavior often can be measured only dichotomously: Students either did or did not enroll in computer-related courses. Therefore, previous research has often used structural equation modeling procedures with dichotomously scaled behavior (e.g., Bagozzi, 1981). Although the path coefficients calculated by LISREL involving dichotomous measures are not maximum likelihood estimates, significant relations between behavioral intentions and behavior would reflect positively on the validity of the intentions measure.

First, a test was performed to determine whether the items designed to measure computer efficacy beliefs and instrumentality beliefs did in fact constitute two distinct constructs. Then the goodness of fit of the entire model was evaluated, and the significance of the path coefficients in the model was assessed.

Results and Discussion

To test whether the items designed to measure computer efficacy beliefs and instrumentality beliefs indeed measure two distinct constructs, both a one-factor and a two-factor model

were fit to these items.¹ The one-factor solution did not fit in either the male or female sample, $\chi^2(18, N = 147) = 50.81, p < .001$, and $\chi^2(18, N = 157) = 58.22, p < .001$, respectively. Next, a two-factor solution was fit to the covariance matrix from the items intended to measure computer efficacy beliefs and instrumentality beliefs. Because there is no a priori reason to assume that these two constructs are orthogonal, the two factors were allowed to correlate but the pattern of factor loadings was fixed so that each item could load only on the factor that it was supposed to measure. This model yielded a good fit to the data in both the male and female samples, $\chi^2(17, N = 147) = 15.33, p > .50$, and $\chi^2(17, N = 157) = 21.81, p > .19$, respectively. Thus, the data support the treatment of computer efficacy beliefs and instrumentality beliefs as separate constructs in the subsequent analyses of the structural relation in the hypothesized model.

Of the 157 female respondents in the original sample, 114 (73%) could be reached by phone 12 weeks later to determine enrollment in classes requiring the use of computers. Of the 147 male respondents, 96 (65%) could be reached.

To test whether sense of efficacy with respect to computers exerts an independent influence on behavioral intentions to purchase or use computers, the a priori model shown in Figure 1 was fit to the data. Using the conventions of causal analysis (e.g., Jöreskog & Sörbom, 1978), Greek letters were used to depict parameters to be estimated, numerals to indicate constrained parameters, circles to represent latent constructs, and boxes to represent measures. In this model, computer efficacy beliefs and instrumentality beliefs each predict behavioral intentions, and behavioral intentions predict behavior. These and all subsequent analyses were based on a sample size of 157 and 147 for the women and men, respectively. All of the analyses were also performed with a sample size of 114 and 96 for women and men, respectively (i.e., the number of respondents available for the 12-week follow-up). Although the chi-square statistics from these analyses were lower (because of the lower sample size) the pattern of results was not affected.

The model presented in Figure 1 yields an overall fit of $\chi^2(49, N = 147) = 64.77, p > .06$, for the male sample, and $\chi^2(49, N = 157) = 63.04, p > .08$, for the female sample. This indicates that the variance or covariance matrices reproduced by the hypothesized model depicted in Figure 1 are (marginally) significantly different from the actual matrices that were empirically obtained from the responses of the male and female samples. In LISREL terminology, the model fits the data only marginally.

However, as Bentler and Bonett (1980) have pointed out, the overall model fit, that is, the comparison of a specified model with the saturated model (a model with 0 *dfs* that would reproduce the covariance matrix perfectly) is often not very informative. The chi-square statistic is a direct function of the number of observations on which the covariance matrix is based, whereas the degrees of freedom are solely dependent on the number of parameters to be estimated in the model. These authors proposed a general normed fit index, Δ , ranging from 0 to 1, where 0 denotes a goodness of fit that is equivalent to that of a model specifying complete independence between variables, and 1 indicates a fit that is equivalent to that of the saturated model (i.e., a model with 0 *dfs* that perfectly reproduces the covariance matrix).

Calculating this fit index yields $\Delta = .88$ for both samples, indicating that the model reproduces most of the covariances among the items. Furthermore, the fit of this model is not significantly worse than that of the less restrictive model in which all latent factors (depicted as circles in Figure 1) are allowed to intercorrelate, that is, the model that includes all possible recursive paths linking latent variables: men, $\chi^2(2, N = 147) = 4.72, ns$, and women, $\chi^2(2, N = 157) = 0.15, ns$. Thus, it can be concluded from these analyses that the hypothesized path model shown in Figure 1 adequately describes the data that were obtained and, furthermore, that additional paths between constructs are not necessary (statistically significant). Table 1 shows the parameter estimates for the standardized solution.

With the exception of one factor loading in the female sample, λ_4 , and the correlation between computer efficacy beliefs and instrumentality beliefs in both samples, all of the parameters are at least twice as large as their respective standard error. The statistical significance of the path coefficients was assessed via incremental chi-square tests (see Bentler & Bonett, 1980). Specifically, the goodness of fit (chi-square associated with the fit) of the model shown in Figure 1 was compared to the fit of a model without a path coefficient linking (a) instrumentality beliefs to intentions, (b) computer efficacy beliefs to behavioral intentions, and (c) intentions to behavior. In effect, these tests assess the significance of the respective path coefficients *after* controlling for all other path coefficients. This procedure thus tests the independent (in a partial correlation sense) contribution of each latent variable.

These analyses showed that behavioral intentions are significantly predicted by instrumentality beliefs—men, $\chi^2(1, N = 147) = 26.50, p < .001$, and women, $\chi^2(1, N = 157) = 27.98, p < .001$ —and computer efficacy beliefs—men, $\chi^2(1, N = 147) = 12.99, p < .001$, and women, $\chi^2(1, N = 157) = 15.34, p < .001$. Furthermore, behavioral intentions predict actual behavior (enrollment in classes requiring the use of computers) 12 weeks after the administration of the questionnaire in both the male and female samples, $\chi^2(1, N = 147) = 48.29, p < .001$, and $\chi^2(1, N = 157) = 29.90, p < .001$, respectively. Excluding any of these parameters from the model in fact always leads to highly significant chi-squares for the overall model fit (all *ps* < .01).

The results of Study 1 show that computer efficacy beliefs make a significant contribution to the prediction of behavioral intentions, independent of beliefs about the instrumental value of learning to use computers. Study 1 also provides evidence for the validity of the behavioral intention measure used. Respondents' actual decisions to enroll in computer science courses 12 weeks after the administration of the questionnaire are significantly predicted by the behavioral intention variable.

Study 2

The purpose of Study 2 was twofold. First, we sought to investigate the role of previous experience with computers in the de-

¹ Because the correlations between Items 1 and 2, and Items 5 and 6, appeared to be greater than their correlation with the other items intended to measure the same construct, their error variances were allowed to correlate.

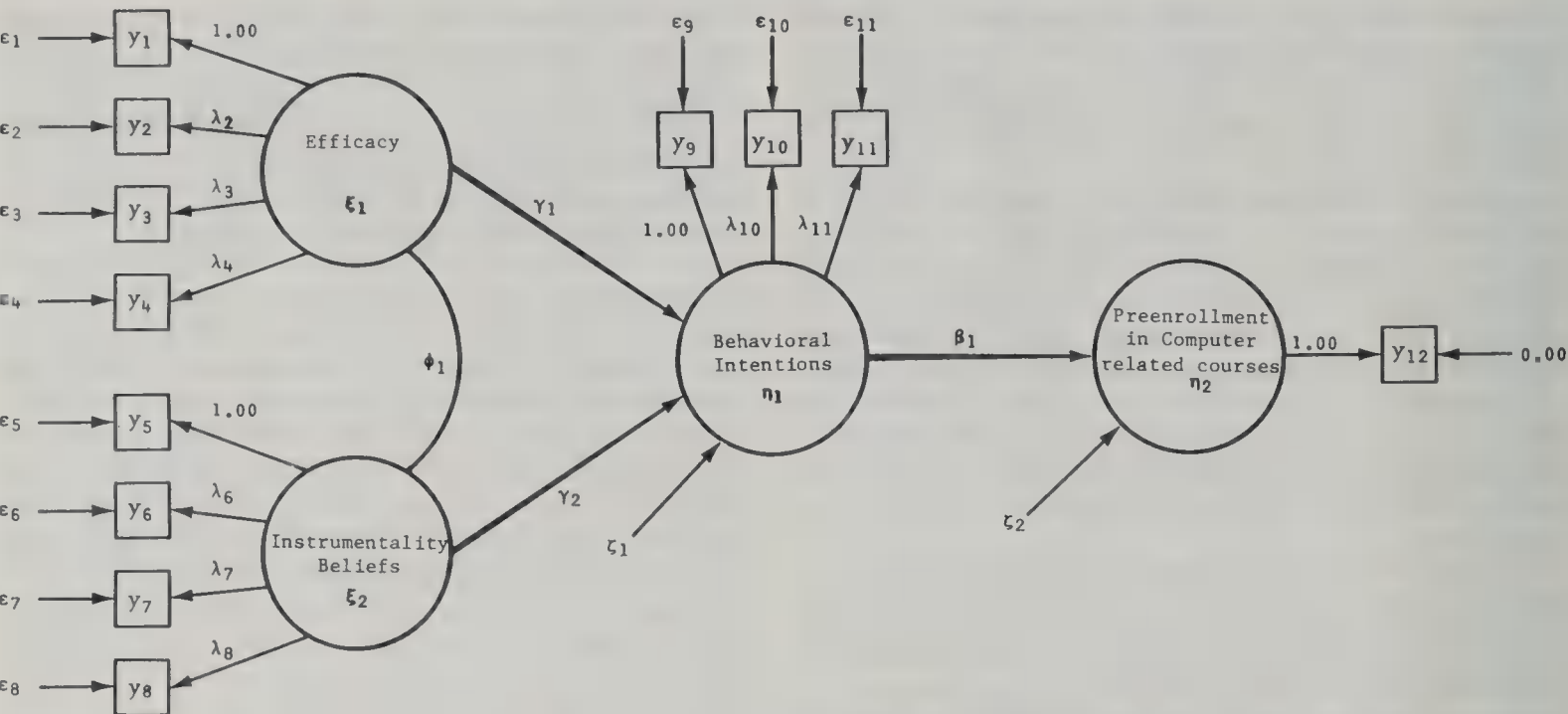


Figure 1. The hypothesized model in Study 1.

cision to adopt computer technology. Previous research by Bandura (1977) on personal efficacy has suggested that direct experience, that is, actually performing a behavior thought to be impossible, is likely to increase one's sense of efficacy and

to reduce phobias. Experience with computers is thus likely to increase personal efficacy with respect to computers. However, experience per se is not likely to exert a direct influence on people's decisions to learn about or use computers, unless computer efficacy beliefs have been affected.

Table 1
Parameter Estimates for Standardized Solution in Study 1

Measure and parameter	Men	Women
Measurement model		
Efficacy beliefs		
λ ₁	.58	.82
λ ₂	.62	.35
λ ₃	.55	.38
λ ₄	.38	.18
Instrumentality beliefs		
λ ₅	.61	.68
λ ₆	.71	.80
λ ₇	.82	.80
λ ₈	.62	.53
Behavioral intentions		
λ ₉	.31	.43
λ ₁₀	.89	.90
λ ₁₁	.76	.83
Behavior		
λ ₁₂	1.00	1.00
Structural model		
β ₁	.59	.45
γ ₁	.43	.35
γ ₂	.53	.56
ψ ₁	.65	.54
ψ ₂	.65	.80
φ ₁	-.24	.06

Note. Estimates are based on the analysis of the correlation matrices.

In Study 2, we used the same items to measure computer efficacy beliefs and instrumentality beliefs as in Study 1. Previous experience with computers, intentions to purchase a micro-computer or to enroll in courses using computers (or both), and pre-enrollment in such courses 2 months later were also assessed. The structural relations between these variables were again tested with linear structural equation modeling procedures (LISREL; Jöreskog & Sörbom, 1978, 1979, 1985).

We predicted that (a) efficacy beliefs would uniquely contribute to the prediction of intentions to purchase or learn about computers, independent of beliefs regarding the instrumental value of using computers, and (b) previous experience with computers would correlate with efficacy beliefs, but would not directly predict intentions to use or learn about computers.

The second purpose of Study 2 was to address the question of whether general beliefs about personal efficacy are related to decisions to use technological innovations in general. Support for this hypothesis would suggest that the theoretical analysis presented here with respect to the adoption of computer technology can also be applied to other technologically advanced products. Use of other electronic innovations was assessed with a set of items adapted from Danko and MacLachlan (1983); general beliefs about personal efficacy were assessed with a scale developed by Paulhus (1983). This scale contains separate subscales for assessing perceived control in the personal and interpersonal spheres. Personal efficacy pertains to control in nonsocial situations (e.g., solving crossword puzzles, building bookcases); interpersonal efficacy relates to control in social situations (e.g., being able to influence other people or to defend

one’s opinions). We predicted that personal efficacy, but not interpersonal efficacy, would correlate with an individual’s use of technologically advanced products.

Method

A questionnaire was administered to a sample of 133 women enrolled in undergraduate psychology courses at a private midwestern university. The questionnaire included the items used in Study 1 to measure efficacy beliefs and instrumentality beliefs regarding computers. In addition, three questions designed to assess previous experience with computers asked respondents how many times in the past year they had used a computer or microcomputer, written a computer program, or used a packaged computer program.

Because Study 2 was conducted at a different university (with different course offerings) than in Study 1, minor adjustments were made to the behavioral intentions scale and the behavioral measure. Five items assessed behavioral intentions to purchase a personal computer, to learn a computer language, to attend any of the seminars offered by the computer center, to take any courses in the following semester that respondents knew would require the use of a computer, or to take a course specifically in computing or data processing. All of the items were accompanied by 10-point scales with appropriate labels. Approximately 8 weeks later, respondents were contacted again and asked whether they had pre-enrolled for the following semester in a computer science course or in any other course requiring the use of a computer.

In addition, Paulhus’s (1983) measures of personal and interpersonal control were administered along with a questionnaire asking subjects to report whether they had ever used any of the following devices (adapted from Danko & MacLachlan, 1983): programmable pocket calculator, automatic garage door opener, automated teller machine, and cordless phone.

Analysis. Analyses of the relations between computer efficacy expect-

tations, instrumentality beliefs, previous experience, behavioral intentions, and subsequent behavior were again performed with LISREL IV and a version of LISREL VI for the IBM/PC (Version VI.9). The covariance matrix was used as input for all analyses. Point-biserial correlation coefficients were calculated to assess the relations between the measures of efficacy in the personal and interpersonal sphere and the use of other technological innovations.

Results

Before an overall model was fit to the data, the unidimensionality of scales used in the structural model was assessed. One-factor models yielded satisfactory fits to the covariance matrices for each construct (all *ps* > .20). As in Study 1, the data support the treatment of computer efficacy beliefs and instrumentality beliefs—as well as previous experience and behavioral intentions—as separate constructs in subsequent analyses of the structural relations in the hypothesized model.

Of the 133 women who completed the questionnaire, 94 (71%) could be reached 8 weeks later. The hypothesized model depicted in Figure 2 was fit to the data. These and all subsequent analyses were based on a sample size of 133. All analyses were also performed with a sample size of 94. Although the chi-square statistics from these analyses were lower (due to the smaller sample size), the pattern of results was not affected.

The overall chi-square associated with this model is $\chi^2(129, N = 133) = 184.77, p < .01$. The general normed fit index (Bentler & Bonett, 1980; see also Results section of Study 1) is $\Delta = .82$, indicating a worse fit than that obtained in Study 1. However, note that the hypothesized model is more constrained than the one tested in Study 1; that is, there are more possibili-

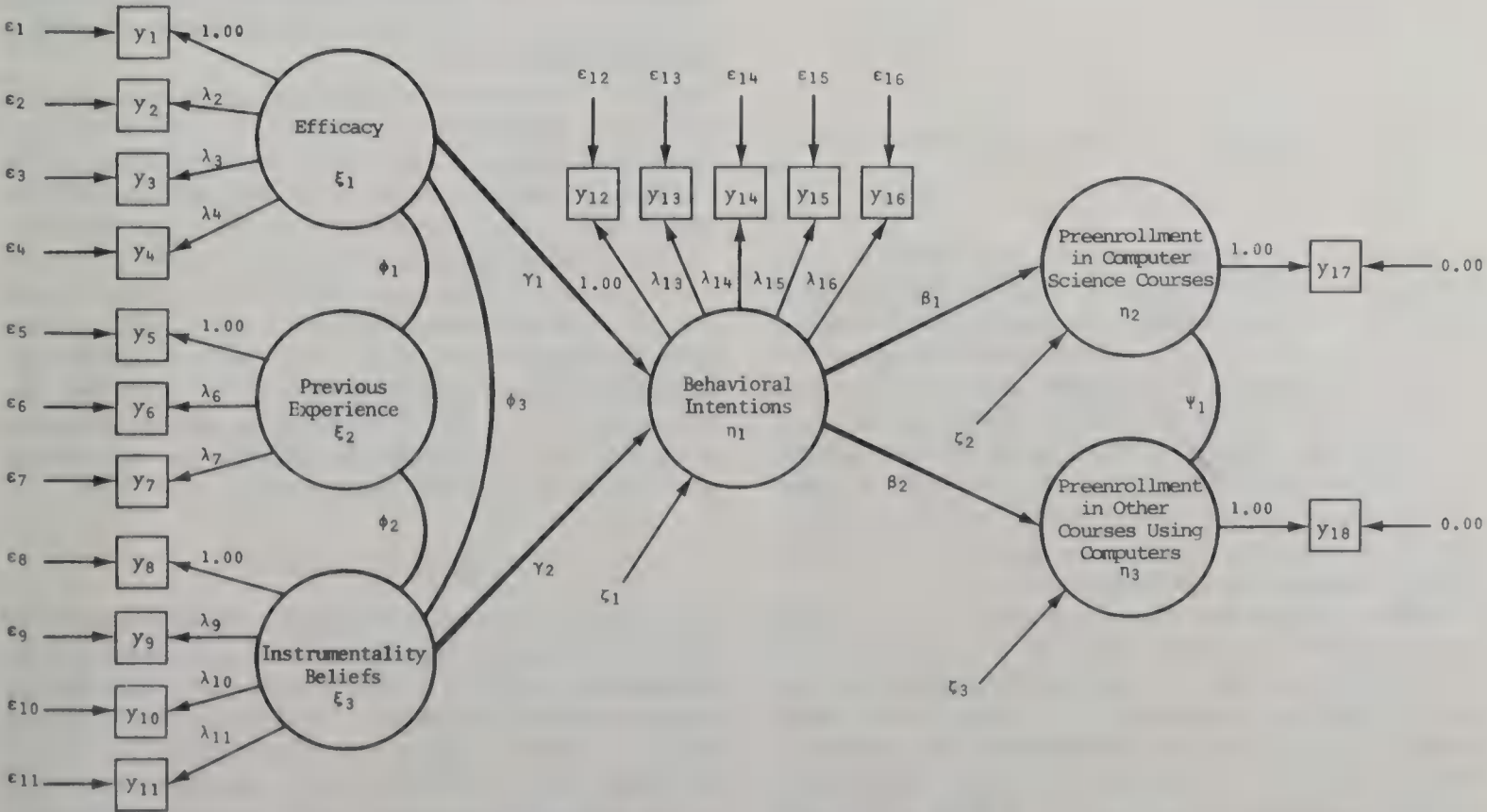


Figure 2. The hypothesized model in Study 2.

Table 2
Parameter Estimates for the Standardized Solution in Study 2

Measure and parameter	Estimate
Measurement model	
Efficacy beliefs	
λ ₁	.74
λ ₂	.66
λ ₃	.50
λ ₄	.41
Previous experience	
λ ₅	1.03
λ ₆	.68
λ ₇	.56
Instrumentality beliefs	
λ ₈	.80
λ ₉	.75
λ ₁₀	.69
λ ₁₁	.66
Behavioral intentions	
λ ₁₂	.48
λ ₁₃	.86
λ ₁₄	.68
λ ₁₅	.75
λ ₁₆	.77
Behavior (η ₁)	
λ ₁₇	1.00
Behavior (η ₂)	
λ ₁₈	1.00
Structural model	
β ₁	.32
β ₂	.27
γ ₁	.59
γ ₂	.39
ψ ₁	.51
ψ ₂	.90
ψ ₃	.93
φ ₁	.48
φ ₂	.19
φ ₃	.01

Note. Estimates are based on the analysis of the correlation matrix.

ties in which the model may not fit the data. (This is reflected in the greater number of degrees of freedom.) Careful inspection of various indices available in LISREL VI (see Jöreskog & Sörbom, 1985) to indicate the location of the lack of fit did not suggest any particular misspecification of the model. Furthermore, the fit of the hypothesized model is not significantly worse than that of a less restrictive model in which all latent variables are allowed to intercorrelate, $\chi^2(7, N = 133) = 4.97, ns$. Thus, it can be concluded from these analyses that the model depicted in Figure 2 adequately describes the data. Table 2 shows the parameter estimates for the standardized solution.

All factor loadings and structural coefficients in the model shown in Figure 2 are greater than twice their standard errors. In addition, incremental chi-square tests show that excluding any of the structural coefficients of the model always significantly decreases the goodness of overall fit. Thus, behavioral intentions significantly predict pre-enrollment both in computer science courses, $\chi^2(1, N = 133) = 8.44, p < .01$, and in other courses requiring the use of computers, $\chi^2(1, N = 133) =$

12.10, $p < .01$. As hypothesized, efficacy beliefs uniquely contribute to the prediction of behavioral intentions, $\chi^2(1, N = 133) = 25.47, p < .01$, as do instrumentality beliefs, $\chi^2(1, N = 133) = 17.41, p < .01$. Subjects who do not believe they could exert control over computers are less inclined to learn about them or to use them. As predicted, previous experience with computers does not significantly contribute to the prediction of behavioral intentions, $\chi^2(1, N = 133) < 1.0$.

Tests of the correlations between latent variables show that in addition to the correlation between the behavioral measures (i.e., enrollment in computer science courses and in other courses requiring the use of computers, $\chi^2[1, N = 133] = 78.53, p < .01$), efficacy beliefs are significantly correlated with previous computer experience, $\chi^2(1, N = 133) = 24.53, p < .01$. No other correlations between latent variables are significant.

A possible alternative to the model shown in Figure 2 is one in which previous experience with computers exerts a direct influence on behavioral intentions, but computer efficacy beliefs do not. Thus, it is conceivable that the effect of efficacy beliefs regarding computers on behavioral intentions are mediated by personal experience. However, this model fits the data very poorly, $\chi^2(129, N = 133), = 210.18, p < .001$, supporting the hypothesis that previous experience with computers is related to computer efficacy beliefs, but that it does not directly predict behavioral intentions to use or learn about computers.

To summarize the relations between latent constructs, we concluded that (a) behavioral intentions to enroll in computer-related courses predict subsequent enrollment, (b) these behavioral intentions are significantly predicted by (related to) beliefs about the instrumental value of learning about computers, (c) behavioral intentions to enroll in computer-related courses are significantly predicted by (related to) computer efficacy beliefs, independent of instrumentality beliefs, and (d) previous experience does not exert a direct influence on intentions to enroll in computer-related courses.

Sphere-specific perceived control and use of other technologies. The reliabilities of the scales used to measure perceived control in the personal and interpersonal spheres were $\alpha = .6$ and $\alpha = .7$, respectively. As predicted, perceived control in the personal sphere was correlated with use of a variety of technologically advanced products: programmable pocket calculators ($r = .22, p < .01$), automated bank teller machines ($r = .16, p < .03$), cordless telephones ($r = .17, p < .03$), and automatic garage door openers ($r = .17, p < .03$). Perceived control in the interpersonal sphere correlated only with use of cordless telephones ($r = .30, p < .001$). It is interesting that this innovation is the only one that is relevant to control in the interpersonal sphere via the facilitation of communication with others.

General Discussion

The results of this research provide evidence that perceived efficacy with respect to computers is an important factor in determining an individual's decision to use them. Moreover, the results regarding sphere-specific measures of perceived control obtained in Study 2 suggest that efficacy beliefs can be sufficiently general to affect an individual's adoption decisions concerning a wide variety of technologically advanced products.

Previous experience with computers does not appear to con-

tribute uniquely to the prediction of behavioral intentions to learn about them. This finding supports the hypothesis that experience per se does not directly affect subsequent behavior regarding further adoption of computer technology; rather, only through changes in perceived efficacy does experience with computer technology lead to a higher likelihood of technology adoption. This finding is consistent with Bandura's (1977) suggestion that *direct* experience of control over a previously avoided task or object is likely to reduce anxieties and induce the individual to change behavior. Future research should be directed at assessing ways to effectively change efficacy beliefs.

References

- Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood Cliffs, NJ: Prentice Hall.
- Bagozzi, R. P. (1981). Attitudes, intentions, and behavior: A test of some key hypotheses. *Journal of Personality and Social Psychology*, 41, 607-627.
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84, 191-215.
- Bandura, A., Adams, N. E., & Beyer, J. (1977). Cognitive processes mediating behavioral change. *Journal of Personality and Social Psychology*, 35, 125-139.
- Bandura, A., & Schunk, D. H. (1981). Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. *Journal of Personality and Social Psychology*, 41, 586-598.
- Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588-606.
- Danko, D. W., & MacLachlan, J. M. (1983). Research to accelerate the diffusion of a new invention: The case of personal computers. *Journal of Advertising Research*, 23, 39-43.
- Dickerson, M. D., & Gentry, J. W. (1983). Characteristics of adopters and non-adopters of home computers. *Journal of Consumer Research*, 10, 225-235.
- Hill, T., & Smith, N. D. (1985, April). *The role of locus of control in the adoption of computer technology*. Paper presented at the meeting of the Southwestern Psychological Association, Austin, TX.
- Hill, T., Smith, N. D., & Mann, M. F. (1986). Communicating innovations: Convincing computer phobics to adopt innovative technologies. In R. J. Lutz (Ed.), *Advances in consumer research*, (Vol. 13, pp. 419-422). Provo, UT: Association for Consumer Research.
- Hirschman, E. C. (1980). Innovativeness, novelty seeking, and consumer creativity. *Journal of Consumer Research*, 7, 283-295.
- Jöreskog, K. G., & Sörbom, D. (1978). *LISREL IV user's guide*. Chicago: National Educational Resources.
- Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt.
- Jöreskog, K. G., & Sörbom, D. (1985). *LISREL VI user's guide*. Morristown, IN: Scientific Software.
- LaBay, D. G., & Kinnear, T. C. (1981). Exploring the consumer decision process in the adoption of solar energy systems. *Journal of Consumer Research*, 8, 271-278.
- Lefcourt, H. M. (1982). *Locus of control: Current trends in theory and research* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Paulhus, D. (1983). Sphere-specific measures of perceived control. *Journal of Personality and Social Psychology*, 44, 1253-1265.
- Phares, E. J. (1978). Locus of control. In H. London & J. E. Exner, Jr. (Eds.), *Dimensions of personality* (pp. 263-303). New York: Wiley.
- Ritchie, E., & Phares, E. J. (1969). Attitude change as a function of internal-external control and communicator status. *Journal of Personality*, 37, 429-443.
- Rogers, E. M. (1962). *Diffusion of innovations*. New York: Free Press.
- Ryckman, R. M., Rodda, W. C., & Sherman, M. F. (1972). Locus of control and expertise relevance as determinants of changes in opinion about student activism. *Journal of Social Psychology*, 88, 107-114.
- Seligman, M. E. P. (1975). *Helplessness: On development, depression, and death*. San Francisco: Freeman.

Received March 12, 1986

Revision received July 24, 1986 ■

Instructions to Authors

Articles submitted for publication in the *Journal of Applied Psychology* are evaluated according to the following criteria: (a) significance of contribution, (b) technical adequacy, (c) appropriateness for the journal, and (d) clarity of presentation. In addition, articles must be clearly written in concise and unambiguous language. They must be logically organized, progressing from a statement of problem or purpose, through analysis of evidence, to conclusions and implications.

Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (3rd ed.). Articles not prepared according to the guidelines of the *Manual* will not be reviewed. All manuscripts must include an abstract of 100–150 words typed on a separate sheet of paper. Typing instructions (all copy must be double-spaced) and instructions on preparing tables, figures, references, metrics, and abstracts appear in the *Manual*. Also, all manuscripts are subject to editing for sexist language.

Authors can refer to recent issues of the journal for approximate length of regular articles. (Three double-spaced manuscript pages equal one printed page.) A few longer articles of special significance are occasionally published as monographs. Short Notes feature brief reports on studies such as those involving some methodological contribution or important replication.

APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more journals. APA policy also prohibits duplicate publication, that is, publication of a manuscript that has already been published in whole or in substantial part in another journal. Also, authors of manuscripts submitted to APA journals are expected to have available their raw data throughout the editorial review process and for at least 5 years after the date of publication.

Authors will be required to state in writing that they have complied with APA ethical standards in the treatment of their sample, human or animal, or to describe the details of treatment. (A copy of the APA Ethical Principles may be obtained from the APA Ethics Office, 1200 17th Street, N.W., Washington, DC 20036.)

Anonymous reviews are optional, and authors who wish anonymous reviews must specifically request them when submitting their manuscripts. Each copy of a manuscript to be anonymously reviewed should include a separate title page with authors' names and affiliations, and these should not appear anywhere else on the manuscript. Footnotes that identify the authors should be typed on a separate page. Authors should make every effort to see that the manuscript itself contains no clues to their identities.

Manuscripts should be submitted in quadruplicate and all the copies should be clear, readable, and on paper of good quality. A dot matrix or unusual typeface is acceptable only if it is clear and legible. Authors should keep a copy of the manuscript to guard against loss. Mail manuscripts to the Editor, Robert Guion, Department of Psychology, Bowling Green State University, Bowling Green, Ohio 43403.

SHORT NOTES

Increasing Voting Behavior by Asking People if They Expect to Vote

Anthony G. Greenwald, Catherine G. Carnot, Rebecca Beach, and Barbara Young
Ohio State University

In two studies, students contacted by telephone were asked to predict whether they would perform a particular behavior (registering to vote or voting, respectively) in the next few days. The proportion who predicted that they would do these socially desirable behaviors exceeded the proportion of control subjects who performed the behavior without first being asked to predict whether they would. Further, in the voting study these errors of overprediction were self-erasing in the sense described by S. J. Sherman (*Journal of Personality and Social Psychology*, 1980, 39). That is, subjects who were asked to predict whether they would vote—all of whom predicted that they would—actually did vote with substantially greater probability than did the no-prediction control subjects. (Actual voting was verified by consulting official voter rolls.) Asking people to predict whether they will perform a socially desirable action appears to increase their probability of performing the action.

When making predictions about their own behavior, people tend to present themselves favorably; they predict that they will do what appears to be proper or good behavior. However, when given the opportunity to act, a person's likelihood of performing a socially desirable action may be reduced by factors such as the action's time and energy costs, the availability of compelling alternatives, and missed opportunity through not responding promptly.

Sherman (1980) showed that asking people to predict their actions does more than just reveal a tendency toward favorable self-presentations; the probability of the predicted action is affected. Once subjects have made a prediction, their behavior is likely to confirm that prediction. In one of Sherman's experiments, subjects who were asked to predict whether they would agree to work 3 hours to collect money for the American Cancer Society (49% said they would) were much more likely (31%) to agree with a later request to do so than were those who were never asked to predict their behavior (4%). Thus, apparent errors in prediction are "self-erasing" (Sherman, 1980). Once a person predicts an action, that action is likely to occur, even when the initial prediction is an apparent gross overestimate of the likelihood of performance. In interpreting this finding, Sherman suggested that making a prediction produces a performance-facilitating cognitive representation in which the person

imagines self-performance of the predicted action and associates that action with supporting reasons.

Sherman's self-erasing-errors-of-prediction finding may be useful as a means of increasing the probabilities of socially desirable actions. The influence technique is remarkably simple: It involves asking people to predict whether they will perform the target action. The present research tested this technique's effectiveness in increasing the probability of performance of two socially desirable behaviors—registering to vote and voting in a national election.

Experiment 1: Voter Registration

Method

Experimenters. The experiment was done as a class project in an honors course in social psychology at Ohio State University in October 1984, a month before the Reagan versus Mondale presidential election. Experimenters were 13 students in the course. Each experimenter was given a complete protocol consisting of (a) an interview script, (b) a set of phone numbers within which to randomly select numbers to be called, and (c) a data sheet on which to record the outcome of each call.

Subjects. Odd-numbered telephone numbers were sampled from the exchange that served the Ohio State University student dormitories. Only students who answered an initial question by reporting that they were *not* registered for the upcoming national election were eligible for inclusion. Students who were registered were asked if they had a roommate who was not registered and who could come to the telephone. There was a high rate of participation among those eligible. However, because only a small proportion of the student population was not registered, only about 15% of answered calls (66 out of 419—131 others were not answered) succeeded in obtaining nonregistered subjects. Later, it was discovered that 4 of the 66 subjects were not properly eligible (3 because they were already registered to vote, and 1 who was not a U.S. citizen). These 4 were dropped from the sample, leaving a sample of 62 subjects.

Procedure. Calls were made on the next to last and last days (Sunday and Monday) before Ohio's registration deadline, which was Tuesday,

The authors are grateful to Steve Hartlage, Megan Varley, Jennifer Martin, Thomas Lah, Maribeth Kuntz, Jeri Lee Ott, Deanna Golden-Kreutz, George Naberezny, Daniel Reed, Karl Rexer, Theresa Jaworski, and Julie Gelpi for their help in the data collection, a class project under the supervision of the first two authors. The third and fourth authors were selected by lottery from among the class members who contributed most to the project, as a means of properly recognizing the major collective contribution of the class members to the research.

Correspondence concerning this article should be addressed to Anthony G. Greenwald, who is now at the Department of Psychology, NI-25, University of Washington, Seattle, Washington 98195.

October 9th, at 7:00 p.m. Callers identified themselves as working on a study of voter knowledge for their social psychology course. When an eligible subject was identified, the caller proceeded to ask the following:

You can help us a lot by answering just a few questions about voter knowledge. I will not be asking for any information about your preferences among candidates or parties. However, because I will be trying to recontact some people before the end of the [term], I will need your name. Are you willing to participate?

Subjects who agreed gave their full names and were asked if they knew, first, where to register to vote and, second, when the registration deadline was. Students who indicated lack of knowledge were given the correct information.

Only after the two information questions were asked were subjects assigned to a treatment by the experimenter's selecting, without replacement, 1 of a set of 10 slips. Each slip was marked either "prediction" or "no prediction." The 30 subjects who were, by this means, assigned to the no-prediction condition were thanked for their help and the phone call was ended. The 32 subjects assigned to the prediction condition were asked an additional question before ending the call:

What do you expect to do between now and the registration deadline of Tuesday evening? Do you expect that you will register to vote or not?

Almost all of the subjects readily answered this question with a yes or no. However, the experimenter was instructed to deal with an "I don't know" response by saying "We would like you to predict your action in any case. Do you think you will register or not?" Those who predicted that they would register were also asked "What would you say is the most important single reason for your registering to vote?" This question was asked on the assumption that providing an explicit reason might increase the probability of subsequently acting in agreement with the prediction (cf. Gregory, Cialdini, & Carpenter, 1982; Sherman, Skove, Hervitz, & Stock, 1981).

Determination of registration. Registration and voting records became available for inspection after the November election. It was expected that almost all of those who registered would register in the election precinct in which their dormitory was located. However, it was possible that some would register instead in their home districts. Follow-up telephone calls were made to all 55 subjects who were not located on the county voter registration rolls. Of the 49 (all but 6) who were successfully recontacted, 16 claimed they had registered in their home locations, rather than in the university area. The remaining 33 confirmed that they had not registered. The 6 who were not recontacted (3 in each condition) were treated as nonregistered. Because it was not possible to verify the responses of the 16 who claimed to be registered outside the university area, the data were analyzed in three ways: (a) treating the 16 "claimants" as if they had not registered, (b) treating them as if they had registered, and (c) dropping them from the sample. (Statistical significance test outcomes were the same for all three analyses.)

Results

Results are summarized in Figure 1 and given in detail in Table 1. As expected, predictions of registration by subjects in the prediction condition (68.8%) significantly exceeded the base rate probability of registration (maximum estimate = 40.6%) by subjects in the no-prediction control condition.

By each of the three methods of determining registration rates in the two conditions, there was about a 10% difference in the expected direction of greater registration in the prediction condition (see Figure 1). However, these differences were not statistically significant.

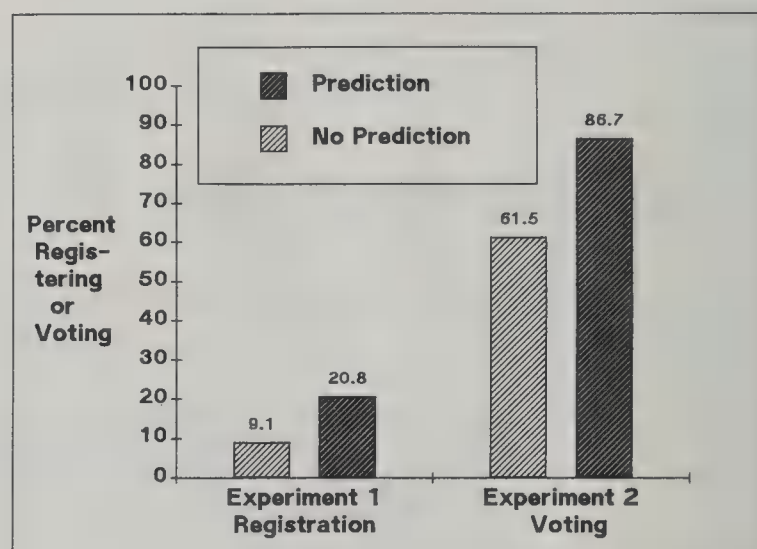


Figure 1. Effects of predicting one's behavior on registering and voting. (This summary uses Method B of Tables 1 and 2, omitting subjects whose claims of having registered or voted could not be confirmed from county records. The results displayed are based on 46 subjects for Experiment 1, and 56 subjects for Experiment 2.)

Experiment 2: Voting

Experiment 2 used approximately the same procedures as Experiment 1 and was conducted before the results of Experiment 1 were known. The study was conducted the Monday evening before the Tuesday, November 6, 1984 election.

Method

Experimenters and subjects. Six undergraduate students, one graduate student, and one faculty member collected data. Eligible participants were resident Ohio State University undergraduates who had registered to vote at the campus election precinct. Experimenters called a total of 452 numbers, 348 of which were answered, and succeeded in contacting a total of 60 students who were eligible and willing to participate. As in Experiment 1, the relatively low yield of subjects was due to the exclusion of the majority of students who were registered at their home addresses rather than at their school addresses. Confining the sample to those registered in the university area was necessary so that county election records could be used as the source of data on voting behavior.

Procedure. Callers randomly sampled only even-numbered telephones within the exchange shared by resident undergraduate students, so as to avoid sample overlap with Experiment 1. Eligible participants were first asked if they knew the location of their voting precinct, and then were asked if they knew the times at which it was open on election day. Most reported that they did know, and those who did not were provided the information. After these two questions, the caller drew a slip to assign the subject randomly to the prediction ($n = 32$) or no-prediction ($n = 28$) condition. For subjects assigned to the prediction condition, before completing the call the experimenter asked, "What do you expect to do between now and the time the polls close tomorrow. Do you expect that you will vote or not?" Parallel to the procedure of Experiment 1, subjects were pressed for an answer to this question, and those who answered that they would vote were asked to provide the most important reason for voting.

Determination of voting. The County Board of Election's voter rolls include the record of whether each registered voter actually voted in the election. In all, 50 of the 60 subjects who reported that they were

Table 1
Registration Behavior in Experiment 1

Treatment	n	Registered in university area	Claimed registration elsewhere	Not registered	Percentage predicting registration	Percentage registering ^a		
						A	B	C
Prediction	32	5	8	19	68.8	15.6	20.8	40.6
No prediction	30	2	8	20	—	6.7	9.1	33.3
Chi-square (1 df)					7.78 ^b	1.24	1.23	0.35

^a These percentages were computed for three different methods of treating the 16 subjects who claimed to have registered outside the university area. Method A treated these subjects as not being registered; Method B dropped them from the sample; and Method C treated them as being registered. The chi-square tests compare the two percentages in each column.

^b This chi-square test compares the percentage predicting registration in the prediction condition with the highest estimate (Method C) of the percentage actually registered in the no-prediction condition. The former is significantly greater at $p < .01$.

registered to vote in the university area were located on the voter registration rolls for the eight voting precincts in the university area. In attempts to follow up by telephone the remaining 10 subjects, 7 were successfully recontacted. Of those 7, 4 claimed they had voted in other locations, and 3 reported that they had not voted. (The 3 who were not contacted—1 in the prediction condition, 2 in the no prediction condition—were classified as nonvoting.) As in Experiment 1, the "claimants" (2 in each condition) were treated in three different ways in analyzing the data, and again, results of significance tests were the same for all three methods.

Results

Table 2 presents details of the results. All 32 (100%) of the subjects in the prediction condition predicted that they would vote. This was highly significantly more than the highest of the three estimates ($18/28 = 64.3\%$) of the percentage of subjects in the no-prediction condition who voted. Again, therefore, the expectation that subjects would overpredict a socially desirable behavior was confirmed.

The percentage of prediction-condition subjects who actually voted was significantly greater by chi-square test ($p < .05$) than the percentage of subjects in the no-prediction condition who voted, for each of the three methods of estimating the percentage who voted. The difference between the two conditions in percentages voting ranged from 23.2% to 25.2% by the three methods (see Figure 1 and Table 2).

Discussion

These two experiments sought to determine whether the phenomenon of self-erasing errors of prediction (Sherman, 1980) could produce consequential effects that are worthy of application. In assessing the results, we consider their application potential, alternately, from the viewpoint of a skeptic and from that of an enthusiast.

A skeptic's first reaction might be to note the limited statistical significance associated with the two findings. The result of the first experiment was simply nonsignificant, and the statistical test of the second experiment exceeded the .05 criterion by only a small margin. Indeed, if the chi-square tests of Experiment 2 were redone using the correction for continuity (see, e.g., Marascuilo & McSweeney, 1977, p. 20) the three alternative tests of the main result—which were reported as significant at $p < .05$ —become results for which the significance level is $.05 < p < .10$. In contrast to this skeptical appraisal, an enthusiast might note that because the direction of result was clearly predicted, a one-tailed statistical test is justified. The result of Experiment 2 is statistically significant at the one-tailed $p < .05$ criterion even when the chi-square correction for continuity is applied.

A skeptic might next note several aspects of the procedures that, although warranted by the circumstances of the present experimental tests, might not characterize an application of the

Table 2
Voting Behavior in Experiment 2

Treatment	n	Voted in university area	Claimed to have voted elsewhere	Not voting	Percentage predicting voting	Percentage voting ^a		
						A	B	C
Prediction	32	26	2	4	100.0	81.3	86.7	87.5
No prediction	28	16	2	10	—	57.1	61.5	64.3
Chi-square (1 df)					13.71 ^b	4.13*	4.69*	4.50*

^a These percentages were computed for three different methods of treating the 4 subjects who claimed to have voted outside the university area. Method A treated these subjects as not having voted; Method B dropped them from the sample; and Method C treated them as having voted. The chi-square tests compare the two percentages in each column.

^b This chi-square test compares the percentage predicting they would vote in the prediction condition with the highest estimate (Method C) of the percentage actually voting in the no-prediction condition. The former is significantly greater at $p < .001$.

* $p < .05$.

phenomenon of self-erasing errors of prediction. Some of these are that (a) the callers (accurately) identified themselves as doing research that was a course project, (b) subjects were asked to give their full names before being asked to predict their behavior, (c) subjects were informed that they might be recontacted later, (d) subjects were sampled from a population that was limited to dormitory-resident college students, and (e) the behaviors studied in both experiments were ones that could be performed in only a narrow time range after the prediction was made. If any of these characteristics constitutes a condition on which the effect of the variation of prediction versus no prediction depends, their absence in another application could undo the effect. In reply to these observations, an enthusiast might note that generalizability of the findings is threatened only on the assumption that one of these factors interacts with the prediction variation to produce the self-erasing-errors-of-prediction phenomenon.¹

Last, a skeptic might observe that the predicted effect was (apparently) obtained in the voting experiment, but not in the registration experiment. Presumably, then, there is some difference between registration and voting behaviors on which the self-erasing-errors-of-prediction phenomenon depends. In response, an enthusiast could note that even the nonsignificant effect of the first experiment was in the predicted direction.

Conclusions

A balanced appraisal may be obtained by considering the magnitude of effects observed in the two experiments. The observed effects were approximately a 10% increase in probability of registration in Experiment 1 and about a 25% increase in probability of voting in Experiment 2. Measured in terms of the w index recommended by Cohen (1977) for describing effect sizes of differences between percentages, the effect in Experiment 1 is approximately $w = .15$, and that for Experiment 2 is approximately $w = .30$. (Cohen, p. 224, identified $w = .10$ and $w = .30$ as "small" and "medium" effects, respectively.) In a large-scale application even the relatively weak effect of Experiment 1 could be of great importance; and the effect observed in Experiment 2 is certainly large enough to alter the outcome of an election. For example, if one could call 10,000 voters who could be counted on to vote for one's preferred candidate, an effect of the strength observed in Experiment 2 would increase that candidate's vote total by about 2,500 votes.

The relative success of Experiment 2 may offer a clue to circumstances under which predicting an action is most likely to increase the rate of performing it. The subjects eligible for Ex-

periments 1 and 2 were, respectively, mutually exclusive subsets of the student population. Subjects in Experiment 1 were among the minority of students who were not registered to vote. Subjects in Experiment 2 were in the majority who were registered. It may have been that registration was a less socially desirable behavior to subjects in Experiment 1 than was voting to subjects in Experiment 2. It is relevant that only 69% of the prediction subjects in Experiment 1 predicted that they would register, in contrast to 100% of the prediction subjects in Experiment 2 predicting that they would vote. Correspondingly, the proportion of control subjects performing the target behavior of registration in Experiment 1 was considerably smaller than the proportion of control subjects in Experiment 2 who performed the target behavior of voting. Another possible difference is that subjects in Experiment 1 may have had less knowledge of how to perform the target behavior of registration than did subjects in Experiment 2 for the target behavior of voting. Thus, it may be that application of the self-erasing-errors-of-prediction finding is more effective the greater the target behavior's social desirability, or the greater the target population's knowledge of how to perform the behavior.

¹ The procedures under which Sherman (1980) obtained self-erasing errors of prediction provide some basis for believing that the effect of the prediction variation is *not* confined to situations in which (a) the procedure is described as research, (b) full names are requested, (c) subjects expect to be recontacted, (d) a dormitory-resident population participates, or (e) the critical actions must be performed in a narrow time range.

References

- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press.
- Gregory, W. L., Cialdini, R. B., & Carpenter, K. M. (1982). Self-relevant scenarios as mediators of likelihood estimates and compliance: Does imagining make it so? *Journal of Personality and Social Psychology*, 43, 89-99.
- Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Belmont, CA: Wadsworth.
- Sherman, S. J. (1980). On the self-erasing nature of errors of prediction. *Journal of Personality and Social Psychology*, 39, 211-221.
- Sherman, S. J., Skov, R. B., Hervitz, E. S., & Stock, C. B. (1981). The effects of explaining hypothetical future events: From possibility to probability to actuality and beyond. *Journal of Experimental Social Psychology*, 17, 142-158.

Received May 1, 1986

Revision received September 17, 1986

Accepted December 1, 1986 ■

Journal of Applied Psychology Monograph

Employee Stock Ownership and Employee Attitudes: A Test of Three Models

Katherine J. Klein
University of Maryland, College Park

Results of a test of three alternative models of the conditions necessary for employee ownership to positively influence employee attitudes are reported. Based on a study of 37 employee stock ownership plan (ESOP) companies (N of individuals = 2,804), results support hypotheses for the extrinsic and instrumental satisfaction models. Average company ESOP satisfaction and organizational commitment are high and average company turnover intention is low when the ESOP provides substantial financial benefits to employees, when management is highly committed to employee ownership, and when the company maintains an extensive ESOP communications program. In contrast, the results provide no support for the intrinsic satisfaction model of ESOP effects.

Employee ownership is increasingly common in American business. In 1976, there were fewer than 1,000 employee ownership companies in the United States (Marsh & McAllister, 1981). Today, there are more than 8,000 (Rosen, Klein, & Young, 1986). Employee ownership is a frequent topic—and a much-touted one—in the popular press (e.g., Hoerr, Stevenson, & Norman, 1985; Posner, 1985). Advocates claim that employee ownership leads to improvements in company performance, productivity, worker participation, and employee morale (e.g., Frieden, 1980; Senate Select Committee on Small Business, 1980). Empirical research on the effects of employee ownership is limited, however. In this article I report the results of the first large-scale study of employee responses to employee ownership: a study of employees' satisfaction with stock ownership, organizational commitment, and turnover intentions in

37 employee stock ownership plan (ESOP) firms. An explication of the basic ESOP structure and a review of previous theory and research on employee ownership precedes the presentation and discussion of the research results.

Employee Stock Ownership Plans

Employee stock ownership plans are deferred employee benefit plans through which employees acquire company stock. Only rarely do employees actually buy the stock themselves, either directly or through wage concessions. Rather, the company gives employees stock as a benefit (in addition to salary and other benefits). More specifically, an ESOP company annually donates stock, or cash to buy stock, to an ESOP trust. Company contributions to the ESOP trust are tax deductible for the company (and remain so under the 1986 revision of the federal tax code). Stock in the trust is allocated to employees' individual accounts, usually based on employee salary. Typically, all full-time employees over the age of 21 are automatically included in the plan after 1 year of service with the company.

Employees' ESOP accounts vest over time. That is, employees earn a gradually increasing right to their allocations, usually 0% for the first 1 to 3 years and increasing regularly to reach 100% after 10 years. Employees receive the vested portion of their ESOP accounts when they leave the company, or in some companies, only when they reach retirement age (even if they had left the company prior to retirement). On receiving their vested ESOP shares, employees in privately held companies may sell their shares back to the company for the current fair market value of the stock. Employees in publicly held ESOP companies may sell their shares on the stock market.

In publicly traded companies, employees must be able to vote their ESOP shares. Privately held companies are required to grant ESOP participants voting rights on a limited number of issues, although they may give employees full voting rights.

Approximately 7,000 U.S. companies have ESOPs. (The re-

This article is based, in part, on material presented in *Employee Ownership in America: The Equity Solution*, by Corey Rosen, Katherine J. Klein, and Karen M. Young (Lexington, MA: Lexington Books, 1986). This material is used with the permission of the authors and of Lexington Books.

The research was supported by National Institute for Mental Health Grant 5R01MH36593 to the National Center for Employee Ownership. I thank Corey Rosen and Karen M. Young for their participation in every aspect of the research project, and Doug Jenkins for his expert advice and support. I also gratefully acknowledge the assistance of Michael Caudell-Feagan, Jane Delgado, Andy Lisak, David Mead-Fox, Allison Read, Valerie Simmons, Sergio Storch, David Toscano, and Bryan Wilson in the collection of the data on several companies. Finally, I thank Tom D'Aunno, Irv Goldstein, John Gomperts, Rosalie Hall, Joan Rentsch, Janice Rouiller, Ben Schneider, Bobbie Turniansky, Trudy Vincent, and two anonymous reviewers for their comments on earlier drafts of this article.

Correspondence concerning this article should be addressed to Katherine J. Klein, Department of Psychology, University of Maryland, College Park, Maryland 20742.

maintaining 1,000 employee ownership companies in the United States are worker cooperatives or have employee stock purchase plans.) Employee stock ownership plans exist in companies of all sizes and industries (ESOP Association, 1982; Marsh & McAllister, 1981). The majority of ESOP companies are privately held and relatively small, with 500 or fewer employees (ESOP Association, 1982; Marsh & McAllister, 1981). The typical ESOP owns between 20% and 40% of company stock, although employees own a majority of company stock in approximately 10% of all ESOPs (Marsh & McAllister, 1981; Rosen et al., 1986). Employee buyouts to rescue failing firms represent only 1 or 2% of all ESOPs (Marsh & McAllister, 1981; Rosen & Klein, 1983). More commonly, companies install ESOPs for one or more of the following reasons: (a) to provide an employee benefit, (b) to provide an incentive for employee effort, (c) to fulfill management's philosophical commitment to shared ownership, (d) to gain tax advantages, (e) to finance capital acquisitions, (f) to purchase the shares of a retiring owner, (g) to finance employee purchase of the company during a corporate divestiture, or (h) to transfer from public to private ownership. More detailed explanations of ESOPs appear in Kaplan and Ludwig (1985) and in Weyher and Knott (1985).

Employee Ownership Theory and Research

The employee ownership literature suggests three alternative, but not mutually exclusive, models of the psychological effects of employee ownership. Underlying each of the three models is the assumption that if employees are satisfied with the employee ownership plan, they will feel committed to the company and motivated to keep working there. Each model predicts that different employee ownership conditions are associated with high employee satisfaction with stock ownership, high organizational commitment, and low turnover intentions.

The first model is the intrinsic satisfaction model of employee ownership. This model suggests that the simple fact of ownership (ownership *qua* ownership) increases employees' commitment to and satisfaction with the company. Tannenbaum (1983), for example, posited that "ownership is attractive to most people. . . . Being an owner is ego enhancing" (p. 251). Long (1978a, 1978b) argued that employee ownership creates a common interest among employees and increases employees' organizational identification. According to this model, the benefits of employee ownership derive directly from ownership, not from a more specific characteristic of the employee ownership plan or of the company as a whole. Thus, the intrinsic satisfaction model is sometimes described as the "direct effects" model of employee ownership (Tannenbaum, 1983).

To test the intrinsic satisfaction model, researchers have typically adopted one of three strategies: (a) examination of the relation between employee attitudes and the number of shares owned by the employee (French & Rosenstein, 1984; Hammer & Stern, 1980), (b) comparison of the attitudes of employee owners and nonowners (Long, 1978a), and (c) comparison of matched employee-owned and conventionally-owned firms (Greenberg, 1980; Rhodes & Steers, 1981; Russell, Hochner, & Perry, 1979). This research has yielded inconsistent, and thus inconclusive, results. In addition, the generalizability of the studies is limited. The studies are all based on small (often sin-

gle-site) samples of worker cooperatives, direct purchase employee buyouts, or employee stock purchase plans. These forms of employee ownership require employees to purchase stock and are often designed to save failing firms. Thus, they have relatively little in common with the vast majority of employee ownership companies in the United States (i.e., ESOPs in profitable firms); the *purchase* of stock in a cooperative, employee buyout, or employee stock purchase plan may create, or reflect, dynamics and expectations very different from the *receipt* of stock as an ESOP benefit. Finally, much of the research suffers methodological problems stemming from the use of employee ownership status or shares as an individual-level independent variable. This measure may be confounded with employee salary, tenure, status, and pre-employee ownership commitment to the company.

The second model is the instrumental satisfaction model of employee ownership. According to this view, employee ownership increases employee influence in company decision making (Stein, 1976), which in turn increases employee commitment. Proponents of this model suggest that employee ownership has a positive impact on employee attitudes if the company provides significant opportunities for worker participation in decision making (Hammer & Stern, 1980; Long, 1978a, 1978b, 1979; Tannenbaum, 1983). This model is sometimes described as the "indirect effects" model of employee ownership (Tannenbaum, 1983).

Research testing the instrumental satisfaction model of employee ownership closely parallels research on the intrinsic satisfaction model. To test the model, researchers have either assessed the relation between ownership shares and employee perceptions of worker influence, or compared employee-owned and nonemployee-owned firms. Indeed, many of the same studies tested both models (Hammer & Stern, 1980; Long, 1978a; Rhodes & Steers, 1981; Russell et al., 1979). As with the first model, the research results are often inconsistent, confounded (again, due to the operationalization of employee ownership), and of limited generalizability (again, due to sample size and employee ownership type). Conte and Tannenbaum's (1978) research examining the relation between the percentage of employee ownership in a company and management's perceptions of worker influence in a sample of 98 employee-owned companies stands out as an exception to this general rule. Unfortunately, this study did not include any measure of employee perceptions.

The final model of employee ownership effects, the extrinsic satisfaction model, suggests that employee ownership increases organizational commitment if employee ownership is financially rewarding to employees. Surprisingly, this model is rarely discussed in employee ownership theory and it has never been tested empirically. Descriptive data on employees' perceptions of the benefits of employee ownership (French & Rosenstein, 1984; Hochner & Granrose, 1985; Long, 1978a; Rosen et al., 1986) lend support to the model, however. The model is further supported by research on pay systems, which documents the importance of financial rewards as a determinant of job satisfaction and organizational choice (Heneman, 1984; Lawler, 1971, 1981). In the language of Katz and Kahn (1978), an ESOP is a "system reward" that may increase employee commitment and decrease employee turnover.

In addition, the extrinsic satisfaction model is consistent with the economic literature on principal and agent relations (Berhold, 1971; Jensen & Meckling, 1976; Lewellen, 1971; Williamson, 1985; Wilson, 1968), which suggests that financial incentives such as merit pay, gainsharing, and—by extension—employee ownership, may make agents' (i.e., employees') utility interests compatible with those of the principal (i.e., the owner). Also potentially relevant is the economics literature on rational expectations (Fischer, 1977; Muth, 1961; Phelps & Taylor, 1977; Taylor, 1983), which examines the extent to which individuals accurately anticipate future macroeconomic trends and—again, by extension—the performance of their own company's stock. A full evaluation and integration of these economic theories with the psychological literature on reward systems and employee ownership is, however, beyond the scope of this article.

In sum, the three employee ownership models provide a useful heuristic for conceptualizing the impact of employee ownership on employees (cf. Elizur, 1984). Unfortunately, employee ownership research testing the models has thus far shed little light on the dynamics and effects of employee ownership, particularly in ESOP firms. As a result, the relevance of the three models to ESOP employee ownership is an open question. In the absence of research and theory on ESOPs *per se*, however, it seems reasonable to turn to the non-ESOP employee ownership literature for preliminary guidelines and hypotheses. Whether the company is a cooperative, direct purchase employee buy-out, or ESOP, the simple fact—the idea—of employee ownership may be important to employees (as the intrinsic satisfaction model suggests). Similarly, regardless of employee ownership form, employees may respond favorably to opportunities to participate in company decision making (instrumental satisfaction) or to the possibility of financial gain (extrinsic satisfaction). Thus, the three models provide an appropriate starting point for ESOP theory building and research.

The Present Study

The present study tests the three employee ownership models by examining the relations between ESOP characteristics and employee attitudes. As such, the study examines ESOP and company characteristics as correlates of company differences in average employee satisfaction with the ESOP, organizational commitment, and turnover intention.

Because the employee ownership models were originally conceived and developed to describe non-ESOP forms of employee ownership, they require some modification in the present context. The intrinsic and extrinsic satisfaction models are readily applicable to ESOP firms. The instrumental satisfaction model requires greater modification. For ease of presentation, I discuss the intrinsic and extrinsic satisfaction models first and then turn to a discussion of the instrumental satisfaction model.

The intrinsic satisfaction model suggests that ownership itself is the critical variable for employee morale in employee ownership companies; the more ownership, the better. According to this model, then, the more company stock that the ESOP owns, the more satisfied employees should be with the ESOP and, to the extent that employees' feelings about the ESOP generalize to positive feelings about the company as a whole, the higher

the employees' organizational commitment and the lower their turnover intentions.

The extrinsic satisfaction model of ESOP effects suggests that the financial benefits of employee ownership are most important for employee satisfaction. Two ESOP characteristics determine how lucrative an ESOP is for the ESOP participants: (a) the size of the company contribution to the ESOP and (b) the return on company stock. The size of the company contribution to the ESOP is calculated as a percentage of employee salary; the larger a company's ESOP contribution, the larger the percentage of salary each company employee acquires through the ESOP. The return on company stock also influences the value of employees' ESOP accounts. If the price of company stock increases, employees' stock accounts are more valuable. Thus, the extrinsic satisfaction model predicts that both the size of the company contribution to the ESOP and stock return are positively related to satisfaction with stock ownership and organizational commitment, and negatively related to turnover intention.

The instrumental satisfaction model posits that employee ownership causes an increase in worker participation, which, in turn, increases employees' organizational commitment and decreases their turnover intentions. The first assumption (that employee ownership causes an increase in worker participation) is inappropriate for the study of ESOP firms, however. Privately held ESOP firms are not required to offer ESOP participants full voting rights. Publicly held firms must offer ESOP participants voting rights, but the ESOP participants are often a small minority among a larger body of public shareholders. Further, there is no evidence to show that ESOP firms are significantly more participative than non-ESOP firms, nor that ESOP firms in which employees own a large percentage of stock are more participative than ESOP firms in which employees own a small percentage of stock.

Nevertheless, the instrumental satisfaction model may have important implications for ESOP firms. The model assumes not only that employee ownership increases worker participation, but also that worker participation increases organizational commitment. This latter assumption is applicable to ESOP firms. If it is correct, then any ESOP characteristic that increases or is associated with employee influence and participation may prove critical for employee satisfaction with the ESOP, as well as with the company as a whole.

Which ESOP characteristics may be expected to increase employee influence? Stock voting rights is the most obvious. Even if voting rights do not guarantee employees much influence (e.g., if employees hold a minority of stock, or if few questions come to a vote), ESOP companies that offer employees voting rights may be more participative than those that do not. Management's decision to offer employees voting rights may reflect a larger commitment on the part of management to involve employees in company decision making through formal or informal mechanisms above and beyond stock voting rights.

Three other ESOP characteristics may also be associated with worker participation and influence. The first is the reason why the company established its ESOP. The assumption here (as before) is that management's reason for ESOP adoption is indicative of, or at least associated with, management style. For example, a company that establishes an ESOP to achieve tax savings

Table 1
Summary of Employee Ownership Models

Model	Predicted cause of employee ownership effects	Relevant ESOP characteristics
Intrinsic satisfaction	Ownership per se	Percentage of company stock owned by ESOP
Extrinsic satisfaction	Financial benefits of ownership	Size of company contribution to ESOP Company stock return
Instrumental satisfaction	Worker participation and influence	ESOP voting rights ESOP reason Management's employee ownership philosophy ESOP communications

Note. ESOP = employee stock ownership plan.

may be less likely to be highly participative than a company that establishes an ESOP to avoid a plant shutdown or to realize management's commitment to worker participation and ownership. Second, management's overall employee ownership philosophy may also be associated with worker participation in company decision making. As before, the more committed that management is to employee ownership as a part of its corporate culture and identity, the more participative the company is likely to be. The final factor is the extent of company communications to employees about the ESOP. An extensive ESOP communications program may also be associated with employee influence and participation insofar as it, like the preceding instrumental satisfaction variables, is indicative of management's larger commitment to employee involvement. In sum, the instrumental satisfaction model predicts that voting rights, ESOP reason, and ESOP communications are all significantly related to the employee outcomes and that the level of worker influence in the company mediates these relations.

The three models thus present three alternative, but complementary, sets of hypotheses about the effects of ESOP employee ownership on employee satisfaction with stock ownership, organizational commitment, and turnover intentions. The three models are summarized in Table 1.

Method

Participants

Data analyses are based on the responses of 2,804 ESOP participants in 37 ESOP companies. The data were collected between May 1982 and November 1984, under the auspices of the National Center for Employee Ownership. Approximately 69% of the respondents were men, and approximately 31% were women. Approximately 91% were White. Almost all of the respondents (92.64%) had completed high school, and nearly 29% had completed college. Average employee tenure was 7.04 years (*SD* = 7.53), and the average respondent was 37.10 years old (*SD* = 11.71). Of the respondents, 34% earned between \$10,000 and \$20,000 a year and 29% earned between \$20,000 and \$30,000 a year. Approximately 13% earned more than \$40,000 a year.

Sampling

At the time of the study, no comprehensive list of ESOP companies was available. Instead, with the assistance of the staff of the National Center for Employee Ownership and student research assistants, I used government and private lists of more than 2,500 ESOP firms to locate potential study companies. Of the approximately 75 ESOP firms that were randomly contacted, 37 agreed to participate in the research. Approximately 10% of the companies that refused to participate said they were undergoing serious financial troubles and they either did not want to be bothered at the time or did not want to survey employees when morale was low. Other companies refused to participate because company management did not think the timing was right for an employee survey, did not think the research was appropriate for their company, or had a policy of refusing all research requests.

Of the participating companies, 22 were located on the west coast, 9 on the east coast, and 6 in other parts of the United States. Nineteen were primarily manufacturing firms, 8 were professional service companies, and 10 were retail or wholesale outfits. All of the companies had established their ESOPs at least 1 year prior to the employee survey. The average ESOP was 6.05 years old (*SD* = 2.85) at the time of the employee survey.

On the basis of existing comparison data (ESOP Association, 1982; Marsh & McAllister, 1981), the company sample appears to be fairly representative of the population of ESOP companies. However, the sample may be biased insofar as (a) some less profitable companies declined to participate, (b) more than one third of the sample companies were majority employee owned, and (c) the sample is composed of firms in which management was sufficiently interested in or concerned about the ESOP to want to participate in the research.

Procedures

A key managerial respondent (the chief executive officer, vice president, or personnel director) in each company was interviewed for background about the company and the ESOP. This individual was typically recommended by the firm for his or her knowledge of the ESOP, as well as for his or her decision-making authority and general knowledge of the firm. The semistructured interview lasted from 1 to 2 hours, and yielded information on ESOP characteristics, basic company characteristics, and management-perceived worker influence. Most of these items (e.g., percentage of company owned by the ESOP, ESOP voting rights, company size) were factual, although several (reason for the ESOP, employee ownership philosophy, and management-perceived worker influence) were potentially subject to individual biases in perception and recollection.

Surveys were distributed to all or—in companies with more than 400 employees—to a random sample of company employees. The average employee response rate in each company was 55.13% (*SD* = 17.15). Employees received a cover letter that explained the purpose of the research and assured employees that their answers were anonymous and their participation voluntary. Employees also received a letter from a company management official encouraging them to answer the survey. In return for participation in the research, each company received feedback on its results.

Measures of Employee Stock Ownership Plan Characteristics

Percentage of company stock owned. Percentage of company stock owned was the number of shares owned by the ESOP relative to the total number of company shares in circulation at the time of the employee survey.

The size of the annual company contribution. The size of the annual

company contribution was the average amount of cash or stock that the company contributed to the ESOP trust in the 3 years preceding the employee survey, expressed as a percentage of the covered employee payroll. In companies that instituted an ESOP less than 3 years before the employee survey, or that failed to provide the contribution data for all 3 years, contribution data for 1 or 2 years were used as appropriate and available. In earlier analyses reported in Rosen et al. (1986), the company contribution for only the single year preceding the employee survey was used. The two contribution measures are correlated .91 ($p < .01$).

Stock return. To measure the performance of company stock, I calculated the 2-year stock return (i.e., the percentage change in the value of company stock during the 2-year period preceding the employee survey). This procedure follows established measures of stock return (Brealey & Myers, 1984), but excludes any measure of stock dividends because none of the sample companies gave out dividends. In privately held ESOP firms ($n = 30$ in the sample), the value of company stock is independently valued once a year. Thus, stock prices fluctuate from year to year, not from month to month or day to day. To assess stock change in publicly held firms ($n = 7$ in the sample), I used the stock price at year end, assessing year to year, not month to month or day to day, changes.

Voting rights. Companies were binary coded to indicate whether the company gave ESOP participants full voting rights.

Reason for the plan. The managerial respondent indicated the primary reason the company established an ESOP from a list of seven reasons: (a) employee benefit, (b) employee incentive, (c) financial and tax purposes, (d) philosophical commitment, (e) avoiding a shutdown, (f) business transfer from existing shareholder(s) to employees, and (g) purchase of the company during a corporate divestiture. The ESOP reason was dummy coded for statistical analyses.

Employee ownership philosophy. To measure management's philosophical commitment to employee ownership, the key managerial respondent answered a scale consisting of three items, each using a 7-point response scale. The three items were "Employee ownership is a central part of our management philosophy," "Employee ownership plays a major role in our corporate culture and identity," and "For our company, the employee ownership plan is primarily a tax-saving or financing mechanism" (reverse scored). The respondent's answers were averaged for the overall employee ownership philosophy score.

Communications about the plan. To measure the extent to which management attempts to inform and educate employees about the ESOP, managerial respondents noted which ESOP communications strategies the company used from a checklist of 12 ESOP communications strategies (e.g., employee ownership discussed in initial orientation for employees, annual meeting for employee stockholders, employee ownership mentioned in company letterhead). The score was a count of how many communications methods the company used.

Measures of Worker Influence

To test the instrumental satisfaction model of ESOP effects, three measures of worker participation and influence were included in the study.

Management-perceived worker influence. The key managerial respondent in each company rated the amount of influence nonmanagerial employees had over seven areas of company decision making (e.g., social events, working conditions, selection of supervisors and management, company financial policy). Managerial respondents rated employee influence over each area on a 5-point scale that ranged from *Workers have no say* (1) to *Workers decide alone* (5). The scale score was the average of the respondent's answers to each of the seven items. This measure was adapted from Kwoka (1976).

Employee-perceived worker influence. Employee respondents completed the same scale as above. Thus, the same items were used to mea-

sure both management-perceived and employee-perceived worker influence. The scale for the key managerial respondent appeared in the management interview, whereas the scale for all employees appeared in the employee survey.

Formal worker participation groups. I used Tannenbaum, Cook, and Lohmann's (1984) scale to measure the formal, established mechanisms for company decision making. The scale was completed by the key managerial respondent at each company. The scale asked whether the company had "any active working groups or committees" pertaining to seven substantive areas (e.g., quality control, strategic planning, budget and financial control). The items were scored to reflect whether each committee included the chief executive officer (1 point), other managerial or supervisory person(s) (2 additional points), and nonsupervisory person(s) (3 additional points), and the items were summed for the total scale score.

Measures of Basic Company Characteristics

To describe the firms and test alternative explanations of the main research results, I collected information on the following basic company characteristics:

Company size. Company size was the number of full-time employees in the company at the time of the employee survey.

Annual sales. Annual sales were the company's total annual sales for the last fiscal year preceding the employee survey.

Unionization. Companies were binary coded to indicate whether any of the company's employees were represented by a union.

Public or private status. Companies were binary coded to indicate whether the company was publicly or privately held.

Shared financial information. During the management interview, the key managerial respondent was asked whether the company shared financial information about company performance (e.g., quarterly and annual reports) with nonmanagerial employees. Companies were binary coded accordingly.

Measures of Employee Attitude Dependent Variables

The three dependent variable measures that follow each used multiple items with a 7-point response scale. Scale scores were created by averaging employee responses to the items in each scale.

Satisfaction with the plan. Eight items measured ESOP participants' satisfaction with stock ownership. Relying on factor analyses, reliability tests, and conceptual analyses, I formed this scale from an original list of 18 items about employee ownership. The original items were adapted from existing reports of the effects of employee ownership (e.g., Conte & Tannenbaum, 1978; Hammer & Stern, 1980) and pilot tested in two firms not included in the present data set. Items in the final scale included "It is very important to me that this company has an employee stock ownership plan," "Owning stock in this company makes me want to stay with this company longer than I would if I did not own stock," and "Owning stock in this company makes me more interested in the company's financial success."

Organizational commitment. The short form (nine positively worded items) of Mowday, Steers, and Porter's (1979) scale measured organizational commitment.

Turnover intention. A three-item scale from the Michigan Organizational Assessment Questionnaire (Cammann, Fichman, Jenkins, & Klesh, 1983) measured turnover intention.

Level of Analysis

The company is the unit of analysis for all of the statistical tests. The dependent variables are the mean company scores on each employee attitude measure (ESOP satisfaction, organizational commitment, and

Table 2

Between-Company Differences in ESOP Satisfaction, Organizational Commitment, and Turnover Intention

Dependent variable	Source	df	MS	F	R ²	Intraclass R
ESOP satisfaction	Between company	36	24.98	20.82*	.22	.21
	Within company	2711	1.20			
Organizational commitment	Between company	36	13.06	12.10*	.14	.13
	Within company	2669	1.08			
Turnover intention	Between company	36	27.06	10.96*	.13	.12
	Within company	2724	2.47			

Note. ESOP = employee stock ownership plan.

* $p < .01$.

turnover intention). The independent variables are global measures describing aspects of the ESOP (e.g., company contribution to the ESOP) or of the company (e.g., company size). I have chosen the company level of analysis over the more common individual level of analysis because my primary research interest is to determine how ESOP companies differ as a function of their unique ESOP and company structures. More specifically, the company-level analyses address the question: How are ESOP and company characteristics related to the average level of employee satisfaction and commitment in the company? These analyses are not intended to address the relation between ESOP or company characteristics and *individual* satisfaction and commitment.

Beyond this conceptual argument, use of the company level of analysis is supported by several specifics of ESOP practice and structure. First, an ESOP is a companywide intervention that management typically establishes in part to influence employees as a group (e.g., to improve overall company morale). As a result, the company level of analysis is most practically relevant to ESOP firms. Second, once an individual is employed by an ESOP firm, participation in the ESOP is not a matter of individual choice. Instead, employees are usually automatically included in the ESOP after 1 year of service with the company. As such, it is rarely appropriate to compare owners and nonowners within an ESOP firm (as might be the case in individual-level analyses). And third, within any single company, the amount of stock owned by the individual is a function of his or her salary and tenure. Thus, tests of the relation between the degree of individual ownership and ESOP satisfaction may effectively be tests of the relation between salary and tenure and ESOP satisfaction. Such analyses fail to test ESOP characteristic effects.

The results of one-way analyses of variance (ANOVAs), presented in Table 2, lend additional support to the choice to use the company level of analysis. The between-company variance in the outcome measures is highly significant, indicating both within-company agreement and between-company variance in ESOP satisfaction, organizational commitment, and turnover intention.

The company-level analyses reported in this study should not be interpreted as representative of individual-level findings (Glick & Roberts, 1984). Analyses at the company level of analysis effectively remove individual-level, within-company variance from consideration. Thus, for example, squared correlations based on the company mean scores show only the proportion of the between-company variance that has been explained, not the proportion of the total within- and between-company variance that has been explained.

Given within-company variance in the dependent measures, the observed relations between ESOP or company characteristics and the aggregated dependent variables are likely to be stronger (greater in magnitude) than the relations one would find between these variables at the individual level of analysis. This reflects the fact that an individual's

score on the dependent variable can be partitioned into two components: (a) the mean company score on the dependent variable (i.e., the between-company component), and (b) unique individual factors (i.e., the within-company component). Not surprisingly, knowing the characteristics of a given ESOP allows one to predict the average company level of ESOP satisfaction more accurately than one can predict the ESOP satisfaction of any specific employee within that company.

In spite of the larger magnitude of the results of company-level analyses, the results of individual-level analyses are often statistically significant at a lower probability level because of the larger sample (e.g., 37 companies vs. 2,804 individuals). These kinds of differences between company-level and individual-level analyses and results should be kept in mind in evaluating the present study results. Finally, it is important to understand that the optimal way to analyze cross-level data remains a matter of continuing debate (Glick, 1980; Glick & Roberts, 1984; Pedhazur, 1982; Rousseau, 1985).

Results

As preliminary background information, Table 3 lists the sample number, Cronbach's alpha, mean, standard deviation, and range of company scores for all continuous variables. Table 4 describes the numerical breakdown for the nominal study variables. A full correlation table of all of the study variables appears in Appendix A.

Table 5 shows the correlations among the employee outcome measures and the ESOP characteristics. (In Tables 5 and 6, r s are reported for all continuous and dichotomous variables and, for ease of comparison, η s is listed for each ANOVA using ESOP reason, the only dummy-coded ESOP characteristic. Analyses of voting rights use the point biserial correlation.) Not surprisingly, the three outcome measures (ESOP satisfaction, organizational commitment, and turnover intention) are highly intercorrelated. Accordingly, one can expect analyses across the three measures to yield consistent results. Conversely, with a few exceptions discussed ahead, the ESOP characteristics are not significantly intercorrelated.

Employee Stock Ownership Plan Characteristics and Employee Outcomes

Table 5 provides a preliminary test of the three employee ownership models. Three of the ESOP characteristics are significantly related to the employee outcomes. Contribution is

Table 3
Number, Reliability, Mean, Standard Deviation, and Range of Continuous Variables

Variable	N	α	M	SD	Range	
					Low	High
ESOP satisfaction	37	.91	4.89	0.63	2.98	5.87
Organizational commitment	37	.90	5.11	0.45	3.92	5.87
Turnover intention	37	.91	2.66	0.63	1.69	4.64
Percentage of company stock owned by ESOP	37	—	42.33	33.11	1.00	100.00
Contribution to ESOP	35	—	9.47	6.28	0.00	25.00
Stock return	33	—	37.05	81.25	−35.00	429.00
Employee ownership philosophy	33	.49	5.14	1.48	1.53	7.00
ESOP communications	33	—	5.58	1.99	1.00	9.00
Management-perceived worker influence	37	.86	2.63	0.65	1.30	4.41
Employee-perceived worker influence	37	.77	2.23	0.30	1.75	3.56
Formal participation groups	34	—	14.59	10.61	0.00	42.00
Company size	37	—	514.00	1167.12	15.00	7,080.00
Company sales	37	—	56 million	165 million	.83 million	1 billion

Note. ESOP = employee stock ownership plan.

positively related to ESOP satisfaction ($r = .50, p < .01$) and organizational commitment ($r = .41, p < .05$), and negatively related to turnover intention ($r = -.50, p < .01$). Management’s employee ownership philosophy shows a similar pattern of results. It too is positively related to ESOP satisfaction ($r = .49, p < .01$) and organizational commitment ($r = .49, p < .01$), and negatively related to turnover intention ($r = -.37, p < .05$). Finally, ESOP communications is positively related to ESOP satisfaction ($r = .40, p < .05$) and negatively related to turnover intention ($r = -.41, p < .05$). These results provide some sup-

port for both the extrinsic satisfaction model and the instrumental satisfaction model. The intrinsic satisfaction model receives no support from the data; percentage of stock owned by the ESOP is not significantly related to the employee outcomes. The remainder of the Results section focuses on the contribution, employee ownership philosophy, and ESOP communications results. Possible reasons for the nonsignificant results for percentage, voting rights, ESOP reason, and stock return are proposed in the Discussion section.

Size of the Company Contribution to the Plan

I turn first to the contribution results and a closer examination of the extrinsic satisfaction model of employee ownership. Does the extrinsic satisfaction model offer the best explanation of the contribution results, or is some other explanation equally or more plausible?

The results in Table 5 indicate that contribution is not significantly related to employee ownership philosophy or ESOP communications. Further, the results in Table 6 indicate that contribution is also not significantly related to any of the measures of basic company characteristics or of worker influence. One can thus conclude that the results in Table 5 linking contribution to the employee outcomes are not explained by any spurious relation between contribution and other ESOP or company characteristics. Partial correlations of contribution and the employee outcomes, controlling for the ESOP and company characteristics listed in Tables 5 and 6, substantiate this assertion. The partial correlations are listed in Appendix B.

Employee Ownership Philosophy and Communications About the Plan

Turning now to the results for employee ownership philosophy and ESOP communications, the instrumental satisfaction model suggests that the two variables should be positively related

Table 4
Description of Nominal Study Variables

Nominal variable	N of companies
ESOP voting rights	
Yes	13
No	24
Unionization	
Yes	9
No	28
Public or private status	
Private	30
Public	7
Shared financial information ^a	
Yes	27
No	7
ESOP reason	
Employee benefit	8
Incentive	4
Financial	4
Philosophical	9
Avoiding a shutdown	1
Business transfer	7
Corporate divestiture	4

Note. ESOP = employee stock ownership plan.
^a Data are missing for three companies.

Table 5
Correlations of Employee Outcomes and ESOP Characteristics

Variable	1	2	3	4	5	6	7	8	9	10
1. ESOP satisfaction	—									
2. Organizational commitment	.87**	—								
3. Turnover intention	-.84**	-.89**	—							
4. Percentage of company stock owned by ESOP	.05	-.13	-.04	—						
5. Voting rights	.16	.15	-.07	-.13	—					
6. ESOP reason	.44	.53	.33	.67**	.35	—				
7. Employee ownership philosophy	.49**	.49**	-.37*	.02	.47**	.62*	—			
8. ESOP communications	.40*	.27	-.41*	.15	.33	.42	.53**	—		
9. Company contribution to the ESOP	.50**	.41*	-.50**	.17	.05	.46	.06	.01	—	
10. Stock return	.23	.29	-.30	-.28	.26	.39	.27	.24	-.08	—

Note. Eta is listed for each one-way analysis of variance using employee stock ownership plan (ESOP) reason. For voting rights, no = 0 and yes = 1.
* $p < .05$. ** $p < .01$.

lated not only to the employee outcomes, but also to the measures of worker influence. The data in Table 6 provide moderate support for this prediction. Management’s employee ownership philosophy is significantly positively related to management-perceived worker influence ($r = .41, p < .05$), and ESOP communications is significantly related to employee-perceived worker influence ($r = .35, p < .05$). Further, the remaining correlations between employee ownership philosophy and ESOP communications and the worker influence measures are in the predicted direction.

The last tenet of the instrumental satisfaction model suggests that worker influence is positively related to employee satisfaction. The results in Table 6 also largely support this hypothesis. Management-perceived worker influence and employee-perceived worker influence are both significantly related to ESOP satisfaction and organizational commitment, and the corre-

lations between these two worker influence variables and turnover intention are in the predicted direction. The measure of formal participation groups is not, however, significantly related to the employee outcomes.

In sum, the correlations in Tables 6 suggest that—as the instrumental satisfaction model predicts—the relations between employee ownership philosophy and ESOP communications and the employee outcomes are at least partially mediated by worker influence. The hierarchical regressions reported in Table 7 provide a direct test of this hypothesis. (The analyses of employee ownership philosophy control for its significant correlate, management-perceived worker influence, whereas the analyses of ESOP communications control for its significant correlate, employee-perceived worker influence.)

The results in Table 7 indicate that after controlling for management-perceived worker influence, employee ownership phi-

Table 6
Correlations of ESOP Characteristics and Employee Outcomes With Company Characteristics and Measures of Worker Influence

Variable	Size	Sales	Union	Private or public	Shared financial information	Management-perceived worker influence	Employee-perceived worker influence	Formal participation groups
Percentage of company stock owned by ESOP	-.14	-.17	-.10	-.41	.20	.23	.19	.06
Voting rights	.26	.26	-.29	.66**	.35*	.41*	.23	.15
ESOP reason	.33	.33	.45	.57	.34	.26	.43	.41
Employee ownership philosophy	.27	.26	-.33	.33	.49**	.41*	.28	.29
ESOP communications	-.01	.00	-.06	.02	.46**	.33	.35*	.26
Company contribution to the ESOP	.11	.09	-.02	-.11	-.28	.18	.08	.05
Stock return	.17	.31	.13	.46**	.10	.03	.00	-.17
ESOP satisfaction	.21	.22	.18	-.02	-.07	.43**	.49**	.22
Organizational commitment	.12	.11	.23	.03	-.09	.39*	.42**	.18
Turnover intention	-.10	-.11	-.34*	.08	.06	-.32	-.26	.01

Note. Eta is listed for each one-way analysis of variance using employee stock ownership plan (ESOP) reason. For voting rights, shared financial information, and unionization, no = 0 and yes = 1.
* $p < .05$. ** $p < .01$.

Table 7
*Hierarchical Regression Analyses of Employee Ownership
 Philosophy and ESOP Communications Controlling for
 Perceived Worker Influence*

Dependent variable	ΔR^2	F
ESOP satisfaction		
Step 1		
Management-perceived worker influence	.17	7.22*
Step 2		
Management's employee ownership philosophy	.12	5.21*
Total	.29	6.21**
Organizational commitment		
Step 1		
Management-perceived worker influence	.14	5.95*
Step 2		
Management's employee ownership philosophy	.14	5.59*
Total	.28	5.77**
Turnover intention		
Step 1		
Management-perceived worker influence	.06	2.32
Step 2		
Management's employee ownership philosophy	.09	3.06
Total	.15	2.69
ESOP satisfaction		
Step 1		
Employee-perceived worker influence	.25	10.67**
Step 2		
ESOP communications	.06	2.62
Total	.31	6.65**
Organizational commitment		
Step 1		
Employee-perceived worker influence	.17	6.18*
Step 2		
ESOP communications	.02	0.68
Total	.19	3.43*
Turnover intention		
Step 1		
Employee-perceived worker influence	.05	1.94
Step 2		
ESOP communications	.13	4.51*
Total	.18	3.22

Note. ESOP = employee stock ownership plan.

* $p < .05$. ** $p < .01$.

losophy remains significantly related to ESOP satisfaction and organizational commitment, although it is no longer significantly related to turnover intention. After controlling for employee-perceived worker influence, ESOP communications is significantly related to turnover intention, but is not significantly related to ESOP satisfaction. (The original correlation of ESOP communications and organizational commitment was not significant.) Thus, worker influence appears to explain much, but not all, of the effects of employee ownership philosophy and ESOP communications on the employee outcomes. (A full list of the partial correlations of both employee ownership philosophy and the employee outcomes, and of ESOP communications and the employee outcomes, controlling for ESOP and company characteristics, appears in Appendix C.)

Combined Effects of Contribution, Employee Ownership Philosophy, and Communications About the Plan

In conclusion, the results for contribution, employee ownership philosophy, and ESOP communications suggest that *both* the extrinsic and instrumental models of ESOP employee ownership are helpful in explaining the conditions under which ESOP employee ownership is associated with positive employee outcomes. Table 8 presents the results of multiple regression analyses examining the overall effects of the significant independent variables (contribution, employee ownership philosophy, and ESOP communications) on the employee outcomes. Because of the strong correlation between employee ownership philosophy and ESOP communications, I used a composite of these two variables (the z score of employee ownership philosophy plus the z score of ESOP communications) for the analyses (Cohen & Cohen, 1983).

The adjusted R^2 results show that together contribution, employee ownership philosophy, and ESOP communications explain between 25% and 39% of the variance in the average company scores for ESOP satisfaction, organizational commitment, and turnover intention.

Discussion

The research results indicate that, on the average, employees are most satisfied with employee ownership and most committed to their companies when the company makes large contributions to the ESOP, when management is highly committed to the concept of employee ownership, and when the company maintains an extensive ESOP communications program. These results support the extrinsic and instrumental satisfaction models of employee ownership. In contrast, the data offer no support for the intrinsic satisfaction model of employee ownership.

Extrinsic Satisfaction Model of Employee Ownership

Contribution. The size of the company contribution to the ESOP is significantly positively related to employee ESOP satisfaction and organizational commitment, and significantly negatively related to turnover intention. The strong effects of ESOP contribution are not explained by any other ESOP or basic company characteristic, nor by the measures of worker influence and participation. Finally, contribution is not subject to common method variance with the dependent variables. The contribution results, thus, strongly suggest that earning a large amount of money through the ESOP leads to positive employee attitudes.

Alternatively, positive employee attitudes might lead to large company contributions to the ESOP *if* management rewarded employee morale by making a large ESOP contribution. Although possible, this chain of events is unlikely. Typically, management is guided by financial, legal, and practical considerations (e.g., the need to achieve tax savings, pay off an ESOP loan, or purchase a retiring owner's stock) in deciding the company's annual contribution to the ESOP.

Much of the literature on financial benefit plans suggests that group- and organization-wide benefit plans have little impact

Table 8

Regressions of Employee Outcomes on Company Contribution to the ESOP and the Composite of Management Employee Ownership Philosophy and ESOP Communications

Dependent variable	Independent variable	β	R^2	Adjusted R^2	F
ESOP satisfaction	Contribution	.42*	.43	.39	11.09*
	Composite of philosophy and communications	.49*			
Organizational commitment	Contribution	.30	.30	.25	6.20*
	Composite of philosophy and communications	.44*			
Turnover intention	Contribution	-.43*	.39	.35	9.30*
	Composite of philosophy and communications	-.44*			

Note. ESOP = employee stock ownership plan.

* $p < .01$.

on individual productivity because the plans do not provide prompt rewards for individual work effort (i.e., the performance-to-outcome expectancy is low for these plans; Heneman, 1984; Lawler, 1981). Although group- and organization-wide plans may have little influence on individual productivity, the results of the present study suggest that the impact of group benefits in general, and ESOPs in particular, on average employee attitudes in a company should not be underestimated.

Finally, given that the size of the company contribution to the ESOP determines the relative benefit of the ESOP for company employees (i.e., the ratio of an employee's annual ESOP benefit to his or her salary), the contribution results suggest that the relative, rather than the absolute, size of an employee benefit may be most important for employee satisfaction with that benefit. Future researchers should examine this relative wealth issue more directly, determining, for example, whether a \$5,000 ESOP account is meaningful and important to employees regardless of employee salary and total income, or whether the \$5,000 ESOP account is more valued by employees with smaller salaries and incomes.

Stock return. Stock return is not significantly related to the employee outcomes. Given the contribution results, the stock return results are somewhat surprising. The most parsimonious explanation for the nonsignificant stock return results is that, compared to the size of the company contribution to the ESOP, the return on company stock has a relatively small influence on the financial rewards that employees receive from the ESOP. It is also possible, however, that a future study, with a larger sample of companies and hence greater statistical power, might find stock return a significant predictor of the employee outcomes; the results are in the predicted direction and they approach statistical significance (e.g., the correlation of stock return and organizational commitment is .29, $p = .10$). Finally, the nonsignificant stock return results may reflect measurement error. Although stock return in privately held firms cannot be measured more precisely than I have measured it here (using stock prices as determined by the company's independent stock valuator), an alternative measure of stock return in publicly held firms,

accounting for daily, weekly, or monthly fluctuations in stock price, could be used in future research.

Instrumental Satisfaction Model of Employee Ownership

Employee ownership philosophy. Employee ownership philosophy, a measure of management's philosophical commitment to employee ownership, is significantly related to the employee outcomes. According to the instrumental satisfaction model, employee ownership philosophy shows this positive relation to the employee outcomes because it is also associated with employee influence, which in turn leads to employee satisfaction. The data largely support this interpretation as management-perceived worker influence does in fact partially mediate the relation between employee ownership philosophy and the employee outcomes.

Still, a few caveats are in order. First, the measure of management's employee ownership philosophy is subjective and amenable to social desirability effects. Second, contrary to predictions, employee ownership philosophy is not significantly related to employee-perceived worker influence nor to formal participation groups. Finally, the direction of causality linking employee ownership philosophy and the employee outcomes is uncertain. Managers may be supportive of employee ownership ideals precisely because their employees show high morale and commitment to the company. The relation between management's employee ownership philosophy and participative management practices deserves attention in future research.

Communications about the plan. ESOP communications is significantly positively related to ESOP satisfaction and significantly negatively related to turnover intention. The results are largely congruent with the instrumental satisfaction model. ESOP communications is significantly related to employee-perceived worker influence, and its relations to the other measures of worker influence are in the predicted direction. In addition, employee-perceived worker influence at least partially mediates the relation between ESOP communications and the employee

outcomes. ESOP communications may also have direct effects on employee attitudes if company communications about the ESOP increase employee understanding of the ESOP and convince employees of management's commitment to the plan.

Stock voting rights. ESOP voting rights is not significantly related to the employee outcomes. This result contradicts the instrumental satisfaction model. Two factors may explain this result. First, as noted earlier, ESOP stock voting rights often provide employees with relatively little power to influence company policies. (Note that voting rights is not significantly related to employee-perceived worker influence.) Second, statistical factors make it difficult to obtain a significant result. Voting rights is a dichotomous variable, whereas the outcome variables are interval scales. Dissimilarities in the shape of the distributions of the independent and dependent variables reduce the maximum possible correlation coefficient (Cohen & Cohen, 1983).

Reason for the plan. ESOP reason is not significantly related to the employee outcomes or to the measures of worker influence and participation. This pattern of results also contradicts the instrumental model of ESOP satisfaction. Apparently, a company's stated reason for establishing an ESOP has little bearing on employee attitudes or management style. Three factors may help to explain the ESOP reason results. First, companies establish ESOPs for several reasons, not just one. Second, the division of ESOP reason into seven specific reasons is somewhat arbitrary. And third, because ESOP reason is dummy coded, it is relatively difficult to obtain significant results with a sample of 37 companies (e.g., a test of the relation of ESOP reason and ESOP satisfaction has *dfs* of 6 and 30).

Intrinsic Satisfaction Model of Employee Ownership

Percentage of employee ownership. The percentage of company stock owned by the ESOP is not significantly related to the average company scores on the employee outcomes. Nor is percentage significantly related to other ESOP characteristics (except ESOP reason) or to measures of worker influence and participation. The percentage results suggest that ESOP employee ownership is not intrinsically rewarding, that there must be an intervening variable—financial gain, participative management, or both—for ESOP employee ownership to be associated with employee satisfaction and commitment.

Conclusion

The contribution results provide strong support for the extrinsic satisfaction model of ESOP employee ownership: money matters. At the same time, the employee ownership philosophy and ESOP communications results suggest the powerful impact of management style on employee attitudes. The research results thus present a balanced picture of ESOP employee ownership. Although perhaps not intrinsically rewarding, ESOP employee ownership does appear to have a positive impact on average employee attitudes when it is coupled with significant financial rewards or participative management practices, or both.

For the growing number of researchers studying employee ownership, the present study answers many questions, but

leaves many unanswered as well: Are the present study results generalizable to other forms of employee ownership—to cooperatives and to employee buyouts to save failing firms? Are the dynamics of employee ownership different when employees actually purchase stock—when stock is *not* a company gift to employees? Is employee satisfaction and commitment higher in employee-owned than in nonemployee-owned firms? Are ESOP companies more participative than non-ESOP firms? Do different kinds of employees (older, higher income, greater tenure) respond differently to employee ownership? These questions deserve attention in future research efforts.

For industrial and organizational psychology and related fields, the study has two broader implications. First, by documenting the impact of financial rewards on employee attitudes, the study invites additional research on compensation and benefit systems. Psychologists have tended to neglect this important area of study. (In contrast, participative management is the focus of considerable psychological research and theory.)

Second, by examining the impact of company level policies and practices on employee attitudes, the study directs psychologists' attention to a new level of analysis and, hence, a new category of research variables. Industrial and organizational psychologists have typically examined the effects of individual-level variables (e.g., job characteristics) on individual attitudes. The present study suggests that it is possible, and indeed valuable, to examine the consequences of company-level factors on employee outcomes. Future research at the company level of analysis may prove practically useful as well, for managers and consultants may ultimately find it more effective to change global, company policies and practices than to address individual-level concerns and complaints.

References

- Berhold, M. (1971). A theory of linear profit sharing incentives. *Quarterly Journal of Economics*, 85, 460–482.
- Brealy, R., & Myers, S. (1984). *Principles of corporate finance* (2nd ed.). New York: McGraw-Hill.
- Cammann, C., Fichman, M., Jenkins, G. D., Jr., & Klesh, J. R. (1983). Assessing the attitudes and perceptions of organizational members. In S. Seashore (Ed.), *Assessing organizational change* (pp. 71–137). New York: Wiley.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). New York: Wiley.
- Conte, M., & Tannenbaum, A. S. (1978). Employee-owned companies: Is the difference measureable? *Monthly Labor Review*, 101, 23–28.
- Elizur, D. (1984). Facets of work values: A structural analysis of work outcomes. *Journal of Applied Psychology*, 69, 379–389.
- ESOP Association. (1982). *ESOP survey, 1982*. Washington, DC: Author.
- Fischer, S. (1977). Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy*, 85, 191–205.
- French, J. L., & Rosenstein, J. (1984). Employee ownership, work attitudes, and power relationships. *Academy of Management Journal*, 27, 861–869.
- Frieden, K. (1980). *Workplace democracy and productivity*. Washington, DC: National Center for Economic Alternatives.
- Glick, W. (1980). Problems in cross-level inferences. In K. H. Roberts & L. Burstein (Eds.), *Issues in aggregation* (pp. 17–30). San Francisco: Jossey-Bass.

- Glick, W. H., & Roberts, K. H. (1984). Hypothesized interdependence, assumed independence. *Academy of Management Review*, 9, 722-735.
- Greenberg, E. S. (1980). Participation in industrial decision-making and worker satisfaction: The case of producer cooperatives. *Social Science Quarterly*, 60, 551-569.
- Hammer, T. H., & Stern, R. N. (1980). Employee ownership: Implications for the organizational distribution of power. *Academy of Management Journal*, 23, 78-100.
- Heneman, R. L. (1984). *Pay for performance: Exploring the merit system* (Work in America Institute Studies in Productivity). New York: Pergamon Press.
- Hochner, A., & Granrose, C. S. (1985). Sources of motivation to choose employee ownership as an alternative to job loss. *Academy of Management Journal*, 28, 860-876.
- Hoerr, J., Stevenson, G., & Norman, J. R. (1985, April). ESOPs: Revolution or ripoff? *Business Week*, pp. 94-108.
- Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3, 305-360.
- Kaplan, J., & Ludwig, R. (1985). *ESOPs*. Washington, DC: Tax Management.
- Katz, D., & Kahn, R. L. (1978). *The social psychology of organizations* (2nd ed.). New York: Wiley.
- Kwoka, J. E., Jr. (1976). The organization of work: A conceptual framework. *Social Science Quarterly*, 57, 632-643.
- Lawler, E. E., III. (1971). *Pay and organizational effectiveness: A psychological view*. New York: McGraw Hill.
- Lawler, E. E., III. (1981). *Pay and organization development*. Reading, MA: Addison-Wesley.
- Lewellen, W. G. (1971). *The ownership income of management*. New York: National Bureau of Economic Research.
- Long, R. J. (1978a). The effects of employee ownership on organizational identification, employee job attitudes, and organizational performance: A tentative framework and empirical findings. *Human Relations*, 31, 29-48.
- Long, R. J. (1978b). The relative effects of share ownership vs. control on job attitudes in an employee-owned company. *Human Relations*, 31, 753-763.
- Long, R. J. (1979). Desires for and patterns of worker participation in decision making after conversion to employee ownership. *Academy of Management Journal*, 22, 611-617.
- Marsh, T. R., & McAllister, D. E. (1981). ESOPs tables: A survey of companies with employee stock ownership plans. *Journal of Corporation Law*, 6, 551-623.
- Mowday, R. T., Steers, R. M., & Porter, L. W. (1979). The measurement of organizational commitment. *Journal of Vocational Behavior*, 14, 224-247.
- Muth, J. (1961). Rational expectations and the theory of price movements. *Econometrica*, 29, 315-333.
- Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction*. New York: Holt, Rinehart, & Winston.
- Phelps, E. S., & Taylor, J. B. (1977). Stabilizing powers of monetary policy under rational expectations. *Journal of Political Economy*, 85, 163-190.
- Posner, B. G. (1985, April). In search of equity. *Inc.*, pp. 51-60.
- Rhodes, S. R., & Steers, R. M. (1981). Conventional vs. worker-owned firms. *Human Relations*, 34, 1013-1035.
- Rosen, C., & Klein, K. J. (1983). Job-creating performance of employee owned firms. *Monthly Labor Review*, 106(8), 15-19.
- Rosen, C., Klein, K. J., & Young, K. M. (1986). *Employee ownership in the United States: The equity solution*. Lexington, MA: Lexington Books.
- Rousseau, D. M. (1985). Issues of level in organizational research: Multi-level and cross-level perspectives. In L. L. Cummings & B. Staw (Eds.), *Research in organizational behavior* (Vol. 7, pp. 1-38). Greenwich, CT: JAI Press.
- Russell, R., Hochner, A., & Perry, S. E. (1979). Participation, influence, and worker ownership. *Industrial Relations*, 18, 330-341.
- Senate Select Committee on Small Business. (1980). *The role of the federal government in employee ownership of business*. Washington, DC: U.S. Government Printing Office.
- Stein, B. A. (1976). Collective ownership, property rights, and control of the corporation. *Journal of Economic Issues*, 10, 298-313.
- Tannenbaum, A. S. (1983). Employee owned companies. In L. L. Cummings & B. Staw (Eds.), *Research in organizational behavior* (Vol. 5, pp. 235-265). Greenwich, CT: JAI Press.
- Tannenbaum, A. S., Cook, H., & Lohmann, J. (1984). *Research report: The relationship of employee ownership to the technological adaptiveness and performance of companies*. Ann Arbor: University of Michigan, Survey Research Center, Institute for Social Research.
- Taylor, J. B. (1983). Rational expectations and the invisible handshake. In J. Tobin (Ed.), *Macroeconomics, prices, and quantities* (pp. 63-82). Washington, DC: Brookings Institute.
- Weyher, H., & Knott, H. (1985). *The employee stock ownership plan* (2nd ed.). Chicago: Commerce Clearinghouse.
- Williamson, O. E. (1985). *The economic institutions of capitalism*. New York: Free Press.
- Wilson, R. (1968). The theory of syndicates. *Econometrica*, 36, 119-132.

Appendix A

Table A-1
Intercorrelations of All Continuous Variables

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. ESOP satisfaction	—																
2. Organization commitment	.87**	—															
3. Turnover intention	-.84**	-.89**	—														
4. Percentage of company stock owned by ESOP	.05	.13	-.04	—													
5. Voting rights	.16	.15	-.07	-.13	—												
6. Management's employee ownership philosophy	.49**	.49**	-.37*	.02	.47**	—											
7. ESOP communications	.40*	.27	-.41*	.15	.33	.53**	—										
8. Company contribution to ESOP	.50**	.41*	-.50**	.17	.05	.06	.01	—									
9. Stock return	.23	.29	-.30	-.28	.26	.27	.24	-.08	—								
10. Size	.21	.12	-.10	-.14	.26	.27	-.01	.11	.17	—							
11. Sales	.22	.11	-.11	-.17	.26	.26	.00	.09	.31	.96**	—						
12. Unionization	.18	.23	-.34*	-.10	-.29	-.33	-.06	-.02	.13	-.04	.02	—					
13. Private or public	-.02	.03	.08	-.41	.66**	.33	.02	-.11	.46**	.42**	.43**	-.11	—				
14. Shared financial information	-.07	-.09	.06	.20	.35*	.49**	.46**	-.28	.10	.16	.13	-.35*	.26	—			
15. Management-perceived worker influence	.43**	.39*	-.32	.23	.41*	.41*	.33	.18	.03	.03	-.05	-.04	.04	.43**	—		
16. Employee-perceived worker influence	.49**	.42**	-.26	.19	.23	.28	.35*	.08	.00	-.12	-.11	.00	-.18	.15	.66**	—	
17. Formal participation groups	.22	.18	.01	.06	.15	.29	.26	.05	-.17	.03	-.00	-.23	-.16	.09	.32	.52**	—

Note. ESOP = Employee stock ownership plan.

* $p < .05$. ** $p < .01$.

Appendix B

Partial Correlations of Contribution to the Employee Stock Ownership Plan (ESOP) and Employee Outcomes

Table B-1 lists the partial correlations between company contribution to the ESOP and ESOP satisfaction, organizational commitment, and turnover intention, controlling separately for each ESOP characteristic, basic company characteristic, and each measure of worker influence.

Table B-1
Partial Correlations of Company Contribution to the ESOP and Employee Outcomes

Control variable	Dependent variable			Control variable	Dependent variable		
	ESOP satisfaction	Organizational commitment	Turnover intention		ESOP satisfaction	Organizational commitment	Turnover intention
Percentage	.50**	.43*	-.51**	Public or private	.50**	.41*	-.50**
Voting rights	.50**	.40*	-.50**	Shared financial information	.50**	.39*	-.50**
Employee ownership philosophy	.47**	.34	-.46**	Management-perceived worker influence	.48**	.37*	-.48**
ESOP communications	.47**	.33	-.49**	Employee-perceived worker influence	.53**	.41*	-.50**
Stock return	.55**	.47**	-.57**	Formal participation groups	.48**	.36*	-.48**
Company size	.49**	.40*	-.50**				
Company sales	.50**	.40*	-.50**				
Unionization	.51**	.43*	-.54**				

Note. ESOP = employee stock ownership plan.
* $p < .05$. ** $p < .01$.

Appendix C

Partial Correlations of Employee Ownership Philosophy and of ESOP Communication with Employee Outcomes

Table C-1 lists the partial correlations between management's employee ownership philosophy and ESOP satisfaction, organizational commitment, and turnover intention, controlling separately for each ESOP characteristic, basic company characteristic, and each measure of worker influence. Table C-2 lists the comparable partial correlations for ESOP communications.

Table C-1
Partial Correlations of Management's Employee Ownership Philosophy and Employee Outcomes

Control variable	Dependent variable		
	ESOP satisfaction	Organizational commitment	Turnover intention
Percentage	.49**	.50**	-.38*
Voting rights	.41*	.44*	-.32
Contribution	.52**	.51**	-.39*
ESOP communications	.35*	.42*	-.20
Stock return	.45*	.36	-.21
Company size	.46**	.48**	-.37*
Company sales	.46**	.48**	-.37*
Unionization	.58**	.62**	-.54**
Public or private	.46**	.46**	-.36*
Shared financial information	.57**	.58**	-.42*
Management-perceived worker influence	.39*	.40*	-.30
Employee-perceived worker influence	.42*	.43*	-.33
Formal participation groups	.45*	.46**	-.39*

Note. ESOP = employee stock ownership plan.
* $p < .05$. ** $p < .01$.

Table C-2
Partial Correlations of ESOP Communications and Employee Outcomes

Control variable	Dependent variable		
	ESOP satisfaction	Organizational commitment	Turnover intention
Percentage	.41*	.30	-.42**
Voting rights	.34	.21	-.37*
Employee ownership philosophy	.19	.01	-.27
Contribution	.44*	.31	-.47**
Stock return	.30	.15	-.33
Company size	.41*	.27	-.41*
Company sales	.41*	.27	-.41*
Unionization	.42*	.29	-.45**
Public or private	.41*	.27	-.41*
Shared financial information	.46**	.32	-.46*
Management-perceived worker influence	.31	.17	-.36*
Employee-perceived worker influence	.28	.15	-.36*
Formal participation groups	.36*	.23	-.42*

* $p < .05$. ** $p < .01$.

Received June 19, 1986
Revision received November 5, 1986
Accepted September 11, 1986 ■



Published quarterly
by the
American Psychological
Association

Mary
Detroit, Michigan 48221
PLEASE DO NOT REMOVE

Volume 72
Number 3

August 1980

Journal of
Applied
Psychology

Editor

Robert M. Gailon

Associate Editors

Irwin L. Goldstein

Frank J. Landy

The *Journal of Applied Psychology* is devoted primarily to original investigations that contribute new knowledge and understanding to any field of applied psychology except clinical psychology. The journal considers quantitative investigations of interest to psychologists doing research or working in such settings as universities, industry, government, urban affairs, police and correctional systems, health and educational institutions, transportation and defense systems, and consumer affairs. A theoretical or review article may be accepted if it represents a special contribution to an applied field.

Editor

Robert M. Guion, *Bowling Green State University*

Associate Editors

Irwin L. Goldstein, *University of Maryland*

Frank J. Landy, *Pennsylvania State University*

Consulting Editors

Lewis E. Albright, *deRecat & Associates, San Francisco, California*

Earl A. Alluisi, *OU SDR, The Pentagon, Washington, DC*

Kenneth M. Alvares, *Frito-Lay, Dallas, Texas*

Phipps Arabie, *University of Illinois*

William B. Askren, *Universal Energy Systems, Dayton, Ohio*

Kathryn M. Bartol, *University of Maryland*

Bernard M. Bass, *State University of New York, Binghamton*

Robert S. Billings, *Ohio State University*

Philip Bobko, *University of Kentucky*

C. Alan Boneau, *George Mason University*

Walter C. Borman, *Personnel Decisions Research Institute,*

Minneapolis, Minnesota

Donald E. Broadbent, *University of Oxford, England*

Wayne F. Cascio, *University of Colorado, Denver*

Margaret M. Clifford, *University of Iowa*

H. Peter Dachler, *Hochschule St. Gallen für Wirts & Sozialwissen, St. Gallen, Switzerland*

Dan R. Dalton, *Indiana University*

Mark L. Davison, *University of Minnesota*

Robyn M. Dawes, *Carnegie-Mellon University*

Fritz Drasgow, *University of Illinois*

Beverly Dugan, *New York Telephone, New York, New York*

E. Ralph Dusek, *Advanced Resource Development Corporation,*
Columbia, Maryland

James L. Farr, *Pennsylvania State University*

Jack M. Feldman, *University of Texas, Arlington*

Jeffrey H. Greenhaus, *Drexel University*

Tove Helland Hammer, *Cornell University*

William C. Howell, *Rice University*

Daniel R. Ilgen, *Michigan State University*

Andrew S. Imada, *University of Southern California*

Lawrence R. James, *Georgia Institute of Technology*

Stanislav V. Kasl, *Yale University*

James G. Kelly, *University of Illinois*

Gary P. Latham, *University of Washington*

Edwin A. Locke, *University of Maryland*

Robert P. Lowman, *Kansas State University*

Ben B. Morgan, Jr., *Old Dominion University*

Karlene H. Roberts, *University of California, Berkeley*

Paul R. Sackett, *University of Illinois, Chicago*

Steven L. Sauter, *NIOSH, Cincinnati*

Frank L. Schmidt, *University of Iowa*

Neal Schmitt, *Michigan State University*

Lyle F. Schoenfeldt, *Texas A&M University*

Stanley E. Seashore, *University of Michigan*

Kirk H. Smith, *Bowling Green State University*

Patricia Cain Smith, *Bowling Green State University*

Barry M. Staw, *University of California, Berkeley*

Mary L. Tenopir, *American Telephone & Telegraph Company, New York*

James R. Terborg, *University of Oregon*

Gary L. Wells, *University of Alberta*

Gary A. Yukl, *State University of New York, Albany*

Sheldon Zedeck, *University of California, Berkeley*

Manuscripts: Submit manuscripts in quadruplicate to the Editor, Robert Guion, Department of Psychology, Bowling Green State University, Bowling Green, OH 43403, according to instructions elsewhere in this journal (see the table of contents). APA and the editors assume no responsibility for statements and opinions advanced by contributors to *Journal of Applied Psychology*.

Change of Address: Send change of address notice and a recent mailing label to the attention of the Subscription Section, APA, 30 days prior to the actual change of address. APA will not replace undelivered copies resulting from address changes; journals will be forwarded only if subscribers notify the local post office in writing that they will guarantee second-class forwarding postage.

Reprints: Authors may order reprints of their articles from the printer when they receive proofs.

Single Issues, Back Issues, and Back Volumes: For information regarding single issues, back issues, or back volumes write to Order Department, APA, 1200 Seventeenth Street, N.W., Washington, DC 20036.

Microform Editions: For information regarding microform editions write to either of the following: University Microfilms, Ann Arbor, MI 48106; or Princeton Microfilms, Princeton, NJ 08540.

Copyright and Permission: Authors must secure from APA and the author of reproduced material written permission to reproduce an article in full or text of more than 500 words. APA normally grants permission contingent upon like permission of the author, inclusion of the APA copyright notice on the first page of reproduced material, and payment of a fee of \$20 per page. Permission from APA and fees are waived for authors who wish to reproduce a single table or figure provided the author's permission is obtained and full credit is given to APA as copyright holder and to the author through a complete citation. Permission and fees are waived for authors who wish to reproduce their own material for personal use; fees only are waived for authors who wish to use more than a single table or figure of their own material commercially. Permission and fees are waived for the photocopying of isolated articles for nonprofit classroom or library reserve use by instructors and educational institutions. Access services may use abstracts without the permission of APA or the author. Libraries are permitted to photocopy beyond the limits of U.S. copyright law; (a) post-1977 articles, provided the per-copy fee in the code for this journal (0021-9010/87/\$00.75) is paid through the Copyright Clearance Center, 27 Congress Street, Salem, MA 01970; (b) pre-1978 articles, provided the per-copy fee stated in the Publishers' Fee List is paid through the Copyright Clearance Center, 27 Congress Street, Salem, MA 01970. Address requests for reprint permission to the Permissions Office, APA, 1200 Seventeenth Street N.W., Washington, DC 20036.

APA Journal Staff: Susan Knapp, *Executive Editor*; Leslie A. Cameron, *Director, Journals Program*; W. Ralph Eubanks, *Manager, Journal Production*; Lois Czapiewski and Theodore J. Baroody, *Production Editors*; Hugh Roberts, *Editorial Intern*; Jodi Ashcraft, *Advertising Sales Manager*.

The *Journal of Applied Psychology* (ISSN 0021-9010) is published quarterly (beginning in February) in one volume per year by the American Psychological Association, Inc., 1400 North Uhle Street, Arlington, VA 22201. Subscriptions are available on a calendar year basis only (January through December). The 1987 rates follow: *Non-member Individual*: \$60 Domestic, \$63 Foreign, \$70 Air Mail. *Institutional*: \$120 Domestic, \$127 Foreign, \$134 Air Mail. *APA Member*: \$30. Printed in the U.S.A. Second-class postage paid at Arlington, VA, and at additional mailing offices. POSTMASTER: Send address changes to the *Journal of Applied Psychology*, 1400 North Uhle Street, Arlington, VA 22201.

-
- 339 Internality and Externality as Correlates of Involvement in Fatal Driving Accidents
I. Montag and Andrew L. Comrey
- 344 Information Requests in the Context of Escalation
Edward J. Conlon and Judi McLean Parks
- 351 Effects of Gain and Loss Decision Frames on Risky Purchase Negotiations
Paul H. Schurr
- 359 Arbitration and Distributive Justice: Equity or Equality?
William W. Notz and Frederick A. Starke
- 366 How Important Are Dispositional Factors as Determinants of Job Satisfaction? Implications for Job Design and Other Personnel Programs
Barry Gerhart
- 374 Unemployment, Job Satisfaction, and Employee Turnover: A Meta-Analytic Test of the Muchinsky Model
Jeanne M. Carsten and Paul E. Spector
- 382 Examination of Avoidable and Unavoidable Turnover
Michael A. Abelson
- 387 Application of Social Learning Theory to Employee Self-Management of Attendance
Colette A. Frayne and Gary P. Latham
- 393 Carpenter Apprentices: Comparison of Career Transitions for Men and Women
Janina C. Latack, Susan L. Josephs, Bonnie L. Roach, and Mitchell D. Levine
- 401 Types and Choices of Performance Feedback
Daniel R. Ilgen and Carol F. Moore
- 407 Effects of Goals and Feedback on Performance in Groups
Tamao Matsui, Takashi Kakuyama, and Mary Lou Uy Onglatco
- 416 Task Complexity as a Moderator of Goal Effects: A Meta-Analysis
Robert E. Wood, Anthony J. Mento, and Edwin A. Locke
- 426 Differences Among Differences: In Search of General Work Preference Dimensions
Robert G. L. Pryor
- 434 Effects of Using High- Versus Low-Performing Job Incumbents as Sources of Job-Analysis Information
Patrick R. Conley and Paul R. Sackett
- 438 Method Variance as an Artifact in Self-Reported Affect and Perceptions at Work: Myth or Significant Problem?
Paul E. Spector
- 444 Situational Leadership Theory: An Examination of a Prescriptive Theory
Robert P. Vecchio

- 452 Effects of Missing Application-Blank Information on Personnel Selection Decisions: Do Privacy Protection Strategies Bias the Outcome?
Dianna L. Stone and Eugene F. Stone
- 457 Stability of Skilled Performance Across Time: Some Generalizations and Limitations on Utilities
Rebecca A. Henry and Charles L. Hulin
- 463 Situational Specificity in Assessment Center Ratings: A Confirmatory Factor Analysis
Peter Bycio, Kenneth M. Alvares, and June Hahn
- 475 Estimating the Standard Error of Projected Dollar Gains in Utility Analysis
Ralph A. Alexander and Murray R. Barrick
- 480 Use of Tests Manifesting Sex Differences as Measures of Intelligence: Implications for Measurement Bias
Mary Roznowski

Short Notes

- 484 Reliability and Validity of the Situational Interview for a Sales Position
Jeff A. Weekley and Joseph A. Gier
- 488 Recognition of Facial Stimuli Following an Intervening Task Involving the Identi-kit
Sara Elizabeth Comish

Monograph

- 493 Meta-Analysis of Assessment Center Validity
Barbara B. Gaugler, Douglas B. Rosenthal, George C. Thornton III, and Cynthia Bentson

Other

- 373 Correction to Earley et al.
- 386 Schmitt Appointed Editor
- 492 Instructions to Authors

Internality and Externality as Correlates of Involvement in Fatal Driving Accidents

I. Montag

Medical Institute of Road Safety and Tel-Aviv University,
Tel-Aviv, Israel

Andrew L. Comrey

University of California, Los Angeles

Previous research has suggested that generalized internality–externality is related to cautious behavior, but attempts to relate Rotter's Internality–Externality (I–E) scale to driving accidents have been disappointing. Scales of internality and externality specifically oriented to driving behavior were developed with the hope that these scales would be more predictive than generalized I–E. These two new scales were administered in Israel to 200 applicants for drivers' licenses and to 200 individuals from the same general population who had been involved in a fatal motor accident. The multiple correlation between the two driving scales and the dichotomous criterion of involvement versus noninvolvement in a fatal driving accident was .38.

It has been suggested in previous research with Rotter's (1966) scale of Internality–Externality (I–E) that externality is related to a lack of caution and a failure to take precautionary steps to avoid the occurrence of unfavorable outcomes (Hoyt, 1973; Phares, 1978; Strickland, 1977, 1978; Williams, 1972).

These findings imply that there should be a relation between this construct and safe driving, despite the fact that some past attempts to relate personality variables to safe driving have not been particularly promising (Knapper & Cropley, 1981; Little, 1970). There have been some encouraging results reported, however (e.g., Cattell, Eber, & Tatsuoka, 1970; Eysenck, 1962; Fine, 1963; Shaw & Sichel, 1971; Signori & Bowman, 1974), so further investigation appears to be warranted in view of the pressing nature of this problem. Preliminary attempts by the first author to relate a translated version of the Rotter (1966) I–E scale to a fatal driving accident criterion resulted in an estimated biserial correlation coefficient of only .12. These disappointing results prompted the exploration of alternative strategies.

Attempts to relate internality–externality to outside criteria have been more successful when the measures of this construct were tailored more specifically to the targeted behavior (e.g., drinking, health, affiliation), rather than using the more general I–E scale itself (Lefcourt, 1981; Wallston & Wallston, 1981; Worell & Tumilty, 1981). A few authors have included items related to driving in studying correlates of internality and externality (Donovan & Marlatt, 1982; Hoyt, 1973; Levenson, 1981).

The first author has developed two specialized scales, a Driving Internality (DI) scale and a Driving Externality (DE) scale, both designed to measure these constructs specifically with reference to driving behavior (see Appendix). Two separate scales

were developed, rather than one bipolar scale, because previous research has shown that the I–E scale itself measures two somewhat negatively correlated dimensions, not a single bipolar dimension (Collins, 1974). As a consequence, Levenson (1981) and Lefcourt (1981), as well as other authors, have developed separate scales for internality and externality. In the present study, it was expected that these two new scales, DI and DE, specifically oriented to driving behavior, would correlate more highly with a criterion of driving safety than the more general Rotter (1966) I–E scale.

Method

Subjects

Subjects for this investigation were 400 psychologically screened, male applicants for drivers' licenses in the state of Israel. In Israel, individuals who are seeking a commercial vehicle license or who have a past history of physical or psychiatric health problems, many traffic violations, or involvement in an accident with a fatality, are required to undergo medical and psychological screening before being licensed.

The 400 subjects involved in this study formed two distinct groups: a random sample of 200 normal cases from the population described, and 200 individuals who had been involved in a fatal driving accident. The normal group had a mean age of 26.5 years ($SD = 6.1$) and 11.0 years of education ($SD = 2.0$). The 200 accident subjects had a mean age of 31.6 years ($SD = 9.2$) and 11.1 years of education ($SD = 2.5$).

Tests Administered

Prior to taking any tests, all of the subjects were given a test of especially transparent lie scale items developed by the first author, the Montag-L scale. Subjects were given their scores on this scale and also feedback about their meaning, in order to reduce any tendency to lie on the subsequently administered tests. Past experience with this procedure has shown that it decreases lying (Montag & Comrey, 1982). This is important with the present subjects, who were motivated to "fake good" in order to be licensed to drive.

Montag Driving Internality and Driving Externality scales. Two scales of 15 six-choice items each were developed by the first author, the

Correspondence concerning this article should be addressed to Andrew L. Comrey, Department of Psychology, University of California, Los Angeles, California 90024.

Table 1
Varimax Rotated Factor Matrix

Item	Scale	DI	DE	HSQ	Item	Scale	DI	DE	HSQ
1	DE	18	49	27	17	DI	62	-14	40
2	DE	-14	42	19	18	DI	35	22	17
3	DE	17	33	14	19	DI	57	01	32
4	DE	01	39	15	20	DI	56	05	32
5	DE	-17	51	29	21	DE	03	47	22
6	DI	58	-21	38	22	DE	30	34	21
7	DI	66	-06	44	23	DE	-11	48	25
8	DI	67	00	44	24	DE	04	35	12
9	DI	62	-05	38	25	DE	-18	38	18
10	DI	61	-20	42	26	DI	45	01	20
11	DE	-06	53	28	27	DI	52	-08	28
12	DE	11	49	26	28	DI	67	-22	50
13	DE	-11	58	35	29	DI	54	-22	34
14	DE	-12	63	41	30	DI	60	-05	36
15	DE	-11	67	46	31	DE	-06	98	97
16	DI	43	05	19	32	DI	99	-12	99

Note. Decimal points have been omitted. Items 31 and 32 are the total DE and DI scale scores, respectively. DI = Driving Internality; DE = Driving Externality; HSQ are communalities.

DI scale and the DE scale. These scales were originally written in Hebrew. English language translations of these items are shown in the Appendix. Items 6 to 10, 16 to 20, and 26 to 30 make up the DI scale, and Items 1 to 5, 11 to 15, and 21 to 25 make up the DE scale. All of the subjects took both the DI and DE scales as part of the screening process to be licensed.¹

Data Analyses Performed

Factor analysis of Driving Internality and Driving Externality items. Pearson product-moment intercorrelations were computed for the 30 DI and DE items and the two total score variables in the sample of 400 cases. This 32 × 32 matrix was factor analyzed by the principal components method, using the minimum residual algorithm developed by Comrey (1973). Only two factors were extracted, inasmuch as only two scales were involved. These two extracted factors were rotated orthogonally by the normal varimax method (Kaiser, 1958).

The addition of the two total scores as variables in this analysis is somewhat unorthodox inasmuch as these variables are linear combinations of the item scores. This reduces the rank of the total correlation matrix by two. Because the main goal here, however, is to see how well the items will be represented by a two-factor solution and how the total DI and DE scores would relate to these two factors, this procedure is considered to be acceptable as a practical means of achieving these limited factor analytic objectives.

T tests of differences between groups. Means were computed for the normal and accident groups for each of the 30 items as well as for the total DI and DE scale scores. *T* tests were carried out to evaluate the statistical significance of these differences.

Multiple correlation. Members of the normal group were given a score of 0, and members of the accident group were given a score of 1. This accident criterion was correlated with the total DI and DE scale scores. Intercorrelations were obtained for the accident criterion variable, the DI scale score, and the DE scale score. Using these correlations, a multiple correlation was obtained between a weighted composite of the DI and DE scale scores and the dichotomous accident criterion variable.

Results

Factor Analysis of Driving Internality and Driving Externality Items

The orthogonal rotated factor loadings are shown in Table 1. Loadings have been rounded to two decimal places and are shown without decimal points.

Factor 1 clearly represents the DI scale, with all DI items loading over .30, and 12 of the 15 items loading above .50. No DI item had a loading over .22 on the DE factor. The DI total score variable had a loading of .99 on Factor 1, the DI factor.

Factor 2, the main DE factor, had loadings above .30 for all 15 DE items, although the loadings generally were not as high as those for the DI items on the DI factor. In all, 9 of the 15 items had loadings above .45. No DE item had a loading of more than .30 on the DI factor. The DE total score variable had a loading of .98 on Factor 2, the DE factor. The factor analytic results shown in Table 1, therefore, lend substantial support to the hypothesis that these 30 items provide good measures of two relatively independent constructs.

Group Differences

Means, standard deviations, differences between means, and *t* ratios are shown in Table 2 comparing the 200 fatal-accident drivers with the 200 normal drivers. These results are given for the 30 DI and DE items and also for the total scale scores.

For the DI items, all differences, except for Item 18, were statistically significant in the expected direction. Item 18 had

¹ A report is being prepared about the development of the Driving Internality and Driving Externality scales, their validity, and other psychometric properties. Included will be correlations with other tests such as the Minnesota Multiphasic Personality Inventory (Hathaway & McKinley, 1967) and the Comrey Personality Scales (Comrey, 1970).

Table 2

Means, Standard Deviations, Differences, and *t* Ratios for DE and DI Items and Total Scores in Two Groups

Item	Accident group		Normal group		<i>M</i> difference	<i>t</i> ratio
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
1	1.67	1.58	1.01	1.29	0.66	4.57***
2	3.71	1.23	2.97	1.53	0.74	5.36***
3	2.82	1.76	2.88	1.85	-0.06	-0.33
4	2.26	1.73	1.68	1.62	0.58	3.43***
5	1.81	1.60	1.01	1.30	0.80	5.47***
6	2.44	1.77	3.13	1.50	-0.69	-4.17***
7	1.90	1.56	2.45	1.51	-0.55	-3.55***
8	2.14	1.58	3.04	1.35	-0.90	-6.11***
9	1.87	1.69	2.94	1.66	-1.06	-6.33***
10	1.15	1.45	2.19	1.65	-1.04	-6.66***
11	1.90	1.55	1.32	1.43	0.58	3.86***
12	1.90	1.48	1.44	1.36	0.46	3.21**
13	3.27	1.27	2.57	1.54	0.70	4.94***
14	2.72	1.47	1.85	1.50	0.88	5.86***
15	2.50	1.53	1.54	1.45	0.96	6.42***
16	2.64	1.53	3.04	1.53	-0.40	-2.60**
17	2.06	1.71	2.77	1.60	-0.72	-4.33***
18	3.50	1.17	3.37	1.28	0.13	1.08
19	3.10	1.38	3.37	1.32	-0.27	-2.01*
20	1.88	1.55	2.45	1.52	-0.56	-3.65***
21	0.86	1.36	0.65	1.14	0.22	1.73
22	2.90	1.53	3.21	1.36	-0.32	-2.18*
23	2.04	1.60	1.67	1.62	0.36	2.24*
24	2.35	1.38	2.49	1.37	-0.14	-0.99
25	2.48	1.65	2.28	1.56	0.20	1.27
26	3.08	1.38	3.42	1.29	-0.34	-2.52*
27	3.48	1.38	3.92	1.10	-0.44	-3.50***
28	1.81	1.62	2.38	1.64	-0.57	-3.46***
29	3.00	1.65	3.63	1.34	-0.64	-4.23***
30	2.33	1.77	2.84	1.60	-0.51	-3.04**
DE	35.08	10.91	28.57	9.64	6.51	6.30***
DI	36.35	13.40	44.94	10.64	-8.59	-7.07***

Note. Data for the accident group are for 200 individuals involved in a fatal accident. Data for the normal group are for 200 persons not involved in a fatal accident. The differences between item and total scale score means are given in the *M* difference column. Table values have been rounded to two decimal places. DE = Driving Externality; DI = Driving Internality.

* $p < .05$. ** $p < .01$. *** $p < .001$.

the smallest loading on the main DI factor. For most of the DE items, differences were statistically significant in the expected direction. Notable exceptions were for Items 3, 22, and 24, which had small mean differences in a direction opposite to that expected.

The DI and DE total scale scores exhibited large differences that were highly significant in the expected direction.

Multiple Correlation

The product-moment correlations of the DI and DE scales with the dichotomous accident criterion were $-.324$ and $.259$, respectively. The correlation between the DI and DE scales was $-.182$. The multiple correlation between DI and DE scales and the dichotomous accident criterion was $.383$.

Discussion

Present findings showing Driving Internality to be negatively related and Driving Externality to be positively related to involvement in fatal accidents are consistent with results of studies cited at the beginning of this article showing that generalized internality rather than externality is related to cautious behavior. Other studies have shown that individuals with an internal locus of control (internals), compared to individuals with an external locus of control (externals), are more attentive and adept at avoiding aversive situations (Lefcourt, 1976; Lefcourt, Gronnerud, & MacDonald, 1973; Strickland, 1977). Still other studies have shown that internals are highly motivated and perform better than do externals in a variety of situations (Lied & Pritchard, 1976; Phares, 1976, 1978; Spector, 1982; Yukl & Latham, 1978).

It can be argued, of course, that individuals involved in a fatal accident would alter their responses on the DI and DE scales to make themselves appear less responsible for the accident. Only a prospective study will provide a definitive answer to this question, and such a study is now in progress. In the meantime, however, some available evidence tends to refute this possible interpretation of the present findings. For example, if accident subjects had raised their externality scores as a result of being involved in an accident, and correspondingly lowered their internality scores, this would have made the correlation between DI and DE scale scores more negative. It was, however, about the same in normal subjects ($-.16$) and accident subjects ($-.18$).

Although the multiple correlation of $.383$ between the DI and DE scales and the dichotomous criterion used in this study is not extremely large, it may be an underestimate of what is possible. Small increments in the *R* might accrue from scale refinements, inasmuch as a few items appear to be functioning marginally. Furthermore, because the normal subjects in this study were from a pool of individuals required to pass screening to be licensed, the *R* could be even higher if a random sample of drivers from the general population had been used as the normal group.

An additional consideration is that the present criterion was a dichotomous one in which there was no attempt to affix blame for the accident. If the criterion could be improved to place in the accident group only individuals adjudged to be substantially at fault, this should raise the *R*. Also, by developing a continuous criterion of accident blame and by using serious accidents that might have been fatal, as well as those that were fatal, greater criterion variance could be obtained that should improve the multiple correlation even further.

On the other side of the coin is the fact that the present study used a 50% base rate for accidents, which is clearly far in excess of the population base rate. This makes it easier to obtain a high correlation. If these results hold up in future investigations, however, and especially if they can be improved on, they may be of substantial assistance in identifying drivers who are at risk. This information, in turn, could be useful in planning interventions that might have an important impact on this serious social problem.

References

- Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the Sixteen Personality Factor Questionnaire (16PF)*. Champaign, IL: Institute for Personality and Ability Testing.
- Collins, B. E. (1974). Four components of the Rotter internal-external scale. *Journal of Personality and Social Psychology*, 29, 381-391.
- Comrey, A. L. (1970). *Manual for the Comrey Personality Scales*. San Diego, CA: Educational & Industrial Testing Service.
- Comrey, A. L. (1973). *A first course in factor analysis*. New York: Academic Press.
- Donovan, D. M., & Marlatt, G. A. (1982). Personality subtypes among driving-while-intoxicated offenders: Relationship to drinking behavior and driving risk. *Journal of Consulting and Clinical Psychology*, 30, 241-249.
- Eysenck, H. J. (1962). The personality of driver and pedestrian. *Medical Science and Law*, 3, 416-423.
- Fine, B. J. (1963). Introversion-extroversion and motor vehicle driver behavior. *Perceptual and Motor Skills*, 12, 95-100.
- Hathaway, S. S., & McKinley, J. C. (1967). *MMPI: Manual for administration and scoring*. New York: Psychological Corporation.
- Hoyt, M. F. (1973). Internal-external control and beliefs about automobile travel. *Journal of Research in Personality*, 7, 288-293.
- Kaiser, H. J. (1958). The varimax criterion for factor analytic rotation in factor analysis. *Psychometrika*, 23, 187-200.
- Knapper, C. K., & Cropley, A. J. (1981). Social and interpersonal factors in driving. In G. M. Stephenson & J. M. Davis (Eds.), *Progress in applied social psychology* (Vol. 1, pp. 191-220). New York: Wiley.
- Lefcourt, H. M. (1976). *Locus of control: Current trends in theory and research*. Hillsdale, NJ: Erlbaum.
- Lefcourt, H. M. (1981). The construction and development of the multidimensional-multiattributational causality scales. In H. M. Lefcourt (Ed.), *Research with the locus of control construct* (Vol. 1, pp. 245-277). New York: Academic Press.
- Lefcourt, H. M., Gronnerud, P., & McDonald, P. (1973). Cognitive activity and hypothesis formation during a double entendre word association test as a function of locus of control and field dependence. *Canadian Journal of Behavioral Science*, 5, 161-173.
- Levenson, H. (1981). Differentiating among internality, powerful others, and chance. In H. M. Lefcourt (Ed.), *Research with the locus of control construct* (Vol. 1, pp. 15-63). New York: Academic Press.
- Lied, T. R., & Pritchard, R. D. (1976). Relationships between personality variables and components of the expectancy-valence model. *Journal of Applied Psychology*, 61, 463-467.
- Little, A. D. (1970). *The state of the art of traffic safety*. New York: Praeger.
- Montag, I., & Comrey, A. L. (1982). Personality construct similarity in Israel and the United States. *Applied Psychological Measurement*, 6, 61-67.
- Phares, E. J. (1976). *Locus of control in personality*. Morristown, NJ: General Learning Press.
- Phares, E. J. (1978). Locus of control. In H. London & J. Exner (Eds.), *Dimensions of personality* (pp. 263-304). New York: Wiley-Interscience.
- Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 80 (1, Whole No. 609).
- Shaw, L., & Sichel, H. (1971). *Accident proneness*. Oxford, England: Pergamon Press.
- Signori, E. K., & Bowman, R. G. (1974). On the study of personality factors in research in driving behavior. *Perceptual and Motor Skills*, 38, 1067-1076.
- Spector, P. E. (1982). Behavior in organizations as a function of employee's locus of control. *Psychological Bulletin*, 91, 482-497.
- Strickland, B. R. (1977). Internal versus external control of reinforcement. In T. Blass (Ed.), *Personality and social behavior* (pp. 219-279). Hillsdale, NJ: Erlbaum.
- Strickland, B. R. (1978). Internal-external expectancies and health-related behaviors. *Journal of Consulting and Clinical Psychology*, 46, 1192-1211.
- Wallston, K. A., & Wallston, B. S. (1981). Health Locus of Control scales. In H. M. Lefcourt (Ed.), *Research with the locus of control construct* (Vol. 1, pp. 189-243). New York: Academic Press.
- Williams, A. F. (1972). Factors associated with seat belt use in families. *Journal of Safety Research*, 4, 133-138.
- Worell, L., & Tumilty, T. N. (1981). The measurement of locus of control among alcoholics. In H. Lefcourt (Ed.), *Research with the locus of control construct* (Vol. 1, pp. 321-333). New York: Academic Press.
- Yukl, G. A., & Latham, G. P. (1978). Interrelationships among employee participation, individual differences, goal difficulty, goal acceptance, goal instrumentality, and performance. *Personnel Psychology*, 31, 305-323.

Appendix

Items for the Driving Internality (DI) and Driving Externality (DE) Scales

No.	Scale	Item stem	No.	Scale	Item stem
1.	DE	Driving with no accidents is mainly a matter of luck	17.	DI	It is always possible to predict what is going to happen on the road and so it is possible to prevent almost any accident
2.	DE	Accidents happen mainly because of different unpredictable events	18.	DI	Accidents happen when the first driver does not take into consideration all the possible actions of the second driver
3.	DE	The driver can do nothing more than drive according to traffic regulations	19.	DI	Accidents happen because the driver does not make enough effort to detect all sources of danger while driving
4.	DE	Accidents happen because of so many reasons we will never know the most important one	20.	DI	Most accidents happen because of lack of knowledge or laziness on the part of the driver
5.	DE	People who drive a lot with no accidents are merely lucky; it is not because they are more careful	21.	DE	If you are to be involved in an accident, it is going to happen anyhow, no matter what you do
6.	DI	The careful driver can prevent any accident	22.	DE	Most accidents happen because the second driver does not pay attention to traffic regulations even when the first driver does
7.	DI	When a driver is involved in an accident, it is because he did not drive as he should	23.	DE	The driver does not have enough control over what happens on the road
8.	DI	When a driver is involved in an accident it is because he did not pay attention to his driving	24.	DE	Most accidents happen because of mechanical failures
9.	DI	Accidents are only the result of mistakes made by the driver	25.	DE	There will always be accidents no matter how much drivers try to prevent them
10.	DI	The driver is to be blamed almost always when an accident occurs	26.	DI	Accidents happen when the driver does not take into consideration all the possible behaviors of pedestrians
11.	DE	It is difficult to prevent accidents in bad conditions such as darkness, rain, narrow roads, curves, and so on	27.	DI	Accident-free driving is a result of the driver's ability to pay attention to what is happening on the roads and sidewalks
12.	DE	Most accidents happen because of bad roads, lack of appropriate signs, and so on	28.	DI	The driver can always predict what is going to happen; that is why there is no room for surprises on the road
13.	DE	It is very hard to prevent accidents involving pedestrians who come out from between parked cars	29.	DI	It is possible to prevent accidents even in the most difficult conditions such as narrow roads, darkness, rain, and so on
14.	DE	Accidents in which children are involved are hard to prevent because they do not know how to be careful	30.	DI	Prevention of accidents depends only on the driver and his characteristics rather than on external factors
15.	DE	It is very hard to prevent accidents in which old people are involved because they cannot hear nor see well			
16.	DI	Accidents happen because drivers have not learned how to drive carefully enough			

Note. These are English translations of the original Hebrew items. Instructions were as follows: You will find in the following some opinions stated by various drivers concerning causes of accidents. Please express your degree of agreement or disagreement with each statement, selecting a number from the following scale: *Disagree very much* (0), *Disagree quite a bit* (1), *Disagree some* (2), *Agree a little* (3), *Agree quite a bit* (4), *Agree very much* (5).

Received July 26, 1985

Revision received January 23, 1987

Accepted January 29, 1987 ■

Information Requests in the Context of Escalation

Edward J. Conlon and Judi McLean Parks

Department of Management Sciences, College of Business Administration, University of Iowa

Staw's (1981) theory of escalation, that decision makers who are responsible for a failure will be more retrospectively oriented than those who are not responsible for a failure, was tested by monitoring the information requests of subjects performing the Adams and Smith decision case (Staw, 1976). A total of 72 Master of Business Administration (MBA) students completed a computer-administered version of the case, in which they were permitted to request information files that had been preclassified as retrospective or prospective on the basis of the results of data collected from a different sample of MBA students. We found that 75% of the subjects who were responsible for a previous failure requested retrospective information, compared to about 25% of the subjects who were not responsible for a failure. This significant difference (i.e., $p < .05$) supported Staw's theory. We also found that the information manipulation eliminated the tendency of subjects who were responsible for failure to escalate allocations.

In the last decade, conditions in which individuals commit themselves to losing courses of action have been examined in several research programs. Researchers have explored a common phenomenon, referred to by different names: escalation (Bazerman, Beekun, & Schoorman, 1982; Bazerman, Guiliano, & Appleman, 1984; Conlon & Wolf, 1980; Staw, 1976; Staw & Fox, 1977), entrapment (Brockner et al., 1984; Brockner, Rubin, & Lang, 1981; Brockner & Rubin, 1985; Rubin & Brockner, 1975), and sunk costs (Arkes & Blumer, 1985; Christensen-Szalanski, & Northcraft, 1985; Northcraft & Wolf, 1984; Thaler, 1980). Although somewhat different experimental methods were used in these programs, all were concerned with behavioral manifestations of continuing a course of action in spite of disappointing results.

Escalation and entrapment effects have been attributed to similar psychological processes. Brockner and Rubin (1985) stated that "decision makers are initially motivated by the desire to have fun, win money or minimize loss of money; they later report that they want to outlast their opponent, or more generally that they have too much invested to quit" (p. 179). Staw (1981) suggested that the goals and information processing of individuals in escalation contexts would be oriented toward rationalizing past actions rather than seeking the most rational further course of action. In both cases, these processes signify a refocusing of attention away from the original reasons for entering a course of action and toward rationalizing previous decisions to enter or remain in the course of action. The object of this study was to test a part of Staw's (1981) theory by examining variation in the information requested by subjects

performing the first experimental task used to demonstrate escalation (Staw, 1976).

Staw (1981) referred to the need to justify past action as *retrospective rationality*. Decision makers are retrospectively rational when, by their decisions, they attempt to justify past events rather than to optimize future outcomes. This need for justification contrasts sharply with *prospective rationality*, the justification of actions based on future outcomes. Because prospective and retrospective rationality are cognitive concepts, indicators of cognitive processing are an appropriate way to identify the extent to which each type of rationality applies to a decision. One indicator suggested by Staw's (1981) theory is information acquisition, which, according to Taylor and Fiske (1981), is a good indicator of cognitive focus. Staw theorized that retrospectively rational individuals will be retrospectively focused and will exhibit greater concern about what happened in the past than will prospectively oriented individuals. Retrospective focusing occurs partly because justification and exoneration require a plausible explanation of how or why a setback occurred. Such explanations are likely to rely on a knowledge of past events. This perspective does not deny that future success, and prospective information, may be important for exoneration, but it does predict that retrospectively rational decision makers will seek information about the past with greater priority than will prospectively rational decision makers.

This study provided the first direct test of Staw's theory about retrospective rationality. Subjects in this study were permitted to request information from a computerized menu prior to making allocation decisions. Based on an a priori scaling, information files were classified as either prospective or retrospective. The following hypothesis predicted the information requests that would result for subjects exposed to the same responsibility and outcome inductions used by Staw (1976).

We gratefully acknowledge the useful comments of Joel Brockner, Jay Christensen-Szalanski, Gary Gaeth, Jerry Rose, and two anonymous reviewers on a draft of this article.

Correspondence concerning this article should be addressed to Edward J. Conlon, Department of Management Sciences, College of Business Administration, University of Iowa, Iowa City, Iowa 52242.

Retrospective information is more likely to be requested prior to prospective information in the search process when the allocator feels responsible for a failure compared to situations of failure for

which the allocator does not feel responsible or any situation of success.

Support for this hypothesis would constitute support for Staw's theory that responsible decision makers are more retrospectively focused following failures than are those who are not responsible for failures.

This study also explored a subsidiary question that arose naturally from the introduction of information search into the experimental methods that produce escalation. The situations that have produced escalation in both the laboratory and in the field are characterized by high degrees of ambiguity. Decision makers in these situations face a context in which the decision alternatives are known, but there are limitations on the extent to which the alternatives can be compared and evaluated. Studies such as this one, in which subjects are allowed to request information in order to study cognitive phenomena, either reduce the level of ambiguity when they provide the requested information or unavoidably sensitize subjects to the lack of information available when they withhold the requested information. Consequently, the typical escalation pattern on allocations may not occur even when the necessary conditions of responsibility and failure are adequately created. In this study we allowed decision makers to request information and then told them that resource limitations would prevent them from seeing the information they requested. Although this tactic would preserve the ambiguity of the situation, it was expected to sensitize allocators to the relative lack of information available prior to their decision. In order to explore the effects of this tactic and to show that the manipulations of responsibility and outcomes were sufficient to produce escalation when information search was not implemented, we contrasted subjects exposed to the usual escalation induction (i.e., the Adams and Smith [A&S] case used by Staw, 1976) with subjects exposed to the same induction with the addition of the information request capability.

Method

Subjects

Subjects were 72 Master of Business Administration (MBA) students at a large midwestern university, who volunteered to participate in a decision-making study in return for extra credit in an organizational theory course. We randomly assigned the subjects to each of the six experimental conditions, an equal number of subjects within each condition.

Procedures

In this study, we used the procedures and most of the materials used by Staw (1976) in his original escalation experiment. The experimental task was the Adams and Smith case, which required the subjects to make two financial allocations to the research and development (R&D) departments of corporate divisions of a manufacturing firm. We created two experimental factors, each having two levels. The manipulation of choice created high-responsibility (i.e., choice) and low-responsibility (i.e., no-choice) conditions. Outcome feedback was manipulated to create a success (i.e., improving performance) and a failure (i.e., continuing declining performance) condition. To examine the effect that the ability to request information prior to the decision would have on allocations, we added two experimental conditions. In those conditions, the subjects

had high responsibility, had no information-request capabilities, and experienced either failure in one condition or success in the other. The result was a $2 \times 2 \times 2$ (Responsibility \times Outcome \times Search) nested design, in which the information-search factor (i.e., the added cells) was nested within the high level of the responsibility factor and fully crossed with the outcome factor. The high-responsibility, failure, no-search condition was comparable to the high-escalation condition found by Staw (1976) in his first demonstration of escalation, making it possible to evaluate the impact of the information-search manipulation on escalation.

The feedback, the descriptions of the corporate divisions, and most of the instructions were identical to those of Staw. The major deviations of our study from the original were the use of a computer to present the experimental materials and collect the data, a limit on the second allocation of \$8 rather than \$20 million, and the information-search aspects of the study. The R&D budget was decreased from \$20 to \$8 million to reflect a more realistic trend of about 11% per year in the rate of budgetary change from the first to the second allocation.

Subjects arrived at the laboratory in groups of up to 12 per session. Each subject was briefed and seated at a computer terminal that had already been logged into a particular experimental condition. The terminals were separated by partitions so that subjects could not communicate with each other during the session. After the students entered their social security numbers into the computer, they were provided with several screens of instructions about the Adams and Smith case (see Staw, 1976, for details).

Responsibility manipulation. In the high-responsibility (i.e., choice) condition, we told subjects that their first task was to select which of two corporate divisions, Industrial Products or Consumer Products, should receive a sum of \$5 million to be budgeted over 5 years and added to the existing R&D budget of the division. Staw's (1976) original study provided sales and earnings data for the past 10 years for each division. For both divisions, the data indicated a steady decline in performance to the point of losses over the previous years, which the board of directors blamed on insufficient R&D. In the present study, we provided the subjects in the search conditions with a numbered information menu. By selecting from the menu, subjects obtained the same sales and earnings data for each division that was provided by Staw, as well as the 5-year R&D plan for each division. Both plans stressed the need to improve manufacturing technologies in order to remain price competitive. All of the subjects in the high-responsibility conditions requested and received both types of information. After their searches were completed, subjects made decisions about which division to fund and wrote a brief memo providing the reasons for their choice. The computer program automatically recorded the information-search protocols of the subjects and the allocation decisions.

In the no-choice condition, we did not require the subjects to make the initial decision. We informed them, as in the original study, that the choice had been made by the previous job holder, and then we showed them the information on which that decision was based.

Outcome manipulation. The manipulation of success and failure was identical to that used in the original study. Depending on which division had received the R&D funds and the experimental condition, we provided the subjects with the same feedback used by Staw (1976), indicating the success or failure of the chosen division following the allocation decision.

Search manipulation. We manipulated information search by either giving or not giving subjects the opportunity to request information prior to making their decisions. In the no-search conditions, which were nested in high responsibility, we gave the subjects the same information available to subjects in the search conditions, prior their making the first decision about which division to fund. On the second decision, about whether to continue funding, we gave subjects in the search conditions

an opportunity to request information, as we will describe. Subjects in the no-search conditions were given no indication that such information was available or that there was any opportunity to request it.

Measurement of information requests. Prior to making their final allocation decisions, the computer offered subjects in the search conditions a choice of five different sources of information on which to base their allocation decisions. Prior to the information search, the program advised the subjects that because of “limitations on time and resources,” they should carefully prioritize their information requests and search the data in order of how important they thought it was to their decision. The first piece of information they selected from the menu should, therefore, be considered the most important by the subjects. Once the subjects had made their first information request and the item requested had been recorded by the computer, they were advised that their time or money—or both—had run out, and that they would have to proceed directly to their allocation decision without the benefit of the requested information. In this way, we were able to ascertain what information the subjects perceived as most relevant, without the content of the information files affecting the actual allocations.

The menu contained five types of information that could be requested by subjects prior to making their allocations. Subjects chose the information on the basis of labels and descriptions that were presented to them on a computerized menu. Two of the menu choices, a 5-year forecast and an R&D prospectus, were designed to be prospectively oriented, that is, oriented toward the future and useful for predicting the future outcomes that would accrue from further allocations. The other three, an R&D report, a set of justification memos, and a CEO (chief executive officer) report were designed to be retrospectively oriented, that is, oriented toward the past and useful for justifying past actions. The menu choices and associated descriptions seen by the subjects are listed in the Appendix. To evaluate this classification, a different sample of 72 MBA students from the same population as the subjects was asked to rate each of the menu choices, which were presented in randomized order exactly as they would appear to the subjects, on four 7-point items indicating the relative prospective or retrospective nature of the information that would be received if each menu item was chosen. The items were designed to evaluate the extent to which the menu text conveyed a past versus future orientation, a usefulness for justifying the past, and a usefulness for determining future outcomes. The ratings on these items were analyzed using a repeated measures multivariate analysis of variance (MANOVA), which simultaneously compared the means obtained on all four items across the five menu choices.

The results indicated that there were significant differences in the ratings across the five menu choices, multivariate $F(16, 55) = 24.5, p < .001$. The items, mean ratings, standard deviations, and 95% confidence intervals are shown in Table 1 for each menu choice. Table 1 indicates that the forecast and R&D prospectus were consistently rated as more future oriented, more useful for deciding how much to allocate, more useful for predicting the effects of additional allocation, and less useful for supporting the past decision to invest when compared to the R&D report, justification memos, and CEO report menu choices. In no case did a 95% confidence interval for the mean ratings of the two prospective choices overlap a mean rating of the three retrospective choices. Based on these results, we concluded that the menu descriptions of the 5-year forecast and the R&D prospectus would be perceived by our subjects to be prospectively oriented, whereas the R&D report, the justification memos and the CEO report would be seen by our subjects as retrospectively oriented.

Allocations. When the subjects indicated that they were ready to make a funding decision, the computer requested that they enter an amount ranging from \$0 to \$8 million to be budgeted to the R&D division of the previously funded division over the next 5 years. The subjects were advised that any funds not allocated would be made available for

Table 1
Mean Ratings of Menu Descriptions on the Pretest Sample (n = 72)

File type	M	SD	95% confidence intervals
Item 1: Future orientation			
5-year forecast	4.82	1.82	4.39–5.24
Prospectus	5.90	1.14	5.63–6.17
R&D report	1.86	1.59	1.58–2.13
CEO report	2.00	1.13	1.73–2.26
Justification summary	2.35	1.48	1.99–2.69
Item 2: Usefulness for determining how much to invest			
5-year forecast	3.68	1.62	3.30–4.07
Prospectus	3.76	1.74	3.36–4.18
R&D report	3.21	1.69	2.82–3.61
CEO report	2.95	1.77	2.53–3.37
Justification summary	2.41	1.48	2.06–2.76
Item 3: Usefulness for supporting past decision to invest (reversed)			
5-year forecast	4.36	1.88	3.91–4.80
Prospectus	5.01	1.63	4.63–5.39
R&D report	2.31	1.47	1.97–2.66
CEO report	2.54	1.28	2.23–2.84
Justification summary	2.77	1.56	2.40–3.14
Item 4: Usefulness for predicting effect of additional investment			
5-year forecast	3.67	1.66	3.27–4.06
Prospectus	3.58	1.77	3.16–4.00
R&D report	3.13	1.71	2.73–3.54
CEO report	2.85	1.77	2.43–3.27
Justification summary	2.27	1.62	1.89–2.66

important capital improvement projects. The computer automatically recorded the amount allocated, thanked the subjects, and logged them off.

Postexperimental measures. Following the computer-mediated portion of the study, we gave the subjects a postexperimental questionnaire. They received two sets of questions designed to evaluate the experimental manipulations. Two questions evaluated the responsibility induction, and two questions evaluated the success or failure induction. The data were collected on 7-point Likert-type scales that ranged from *not at all* (1) to *a very great extent* (7). The questions are shown in Table 2.

Results

Manipulation Checks

To evaluate the effectiveness of the manipulations of responsibility and outcome, we conducted a MANOVA on each pair of manipulation checks. The means for the manipulation checks items are displayed by condition in Table 2. The analysis showed that the choice manipulation produced significant main effects in the expected directions on the responsibility items, multivariate $F(4, 63) = 6.18, p < .001$. There were no significant interaction effects. Similarly, the success or failure manipulation produced a significant main effect on the outcome items, multivariate $F(4, 63) = 18.34, p < .001$, in the expected direction. Again, there were no significant interactions.

Table 2
Results by Experimental Condition on Manipulation Checks, Information Requests, and Allocations

Manipulation check	Low-responsibility search		High-responsibility			
			Search		No search	
	Success	Failure	Success	Failure	Success	Failure
Outcome A ^a						
<i>M</i>	4.08	2.83	5.67	2.66	5.08	2.33
<i>SD</i>	1.83	2.03	1.07	1.50	1.56	1.61
Outcome B ^b						
<i>M</i>	4.83	2.50	6.00	1.58	5.00	2.08
<i>SD</i>	1.64	1.78	0.60	1.16	1.80	1.68
Responsibility A ^c						
<i>M</i>	2.67	3.33	5.50	5.75	4.83	5.17
<i>SD</i>	2.31	2.35	1.00	0.62	2.48	1.99
Responsibility B ^d						
<i>M</i>	3.00	2.83	4.67	3.92	4.58	3.58
<i>SD</i>	2.09	1.64	1.50	1.83	2.23	2.39
Allocation (in millions of dollars)						
<i>M</i>	3.90	3.39	4.58	3.08	3.88	5.33
<i>SD</i>	2.47	2.44	2.11	3.03	2.11	3.51
Information requests						
Prospective	10	9	10	3	—	—
Retrospective	2	3	2	9	—	—

^a To what extent would you regard the effect of funding as a success? ^b To what extent did the research and development funding produce the desired result? ^c To what extent did you feel responsible for the choice of which division to fund? ^d To what extent did you feel responsible for the performance of the division that was funded?

Information Requests

The hypothesis, which predicted patterns of information requests, was tested using a 2 × 2 (Responsibility × Outcome) design. Because information requests could not be made in the no-search conditions, the no-search cells were excluded from the analysis of information requests without affecting the estimated error or biasing the results. The appropriate interaction for testing the hypothesis contrasted the high-responsibility failure condition with the pooled results of all other cells.

To test the hypothesis that retrospectively relevant information would be requested following a high-responsibility failure, we constructed a 0–1 measure of information search by assigning a 0 to those subjects who requested retrospective information (i.e., the R&D report, justification memos, and CEO report) and a 1 to those requesting prospective information (i.e., the 5-year forecast and future R&D prospectus). The distribution of 0s and 1s across experimental conditions is shown in Table 2. Using binomial logit analysis (Agresti, 1984), we examined the effects of the responsibility and outcome manipulations on the information-search measure. The contrast that compared the high-responsibility failure condition with all others was the only significant effect ($z = 2.99, p < .01$). This interaction effect was the appropriate test of the hypothesis. It occurred because subjects in the high-responsibility failure condition were predominantly focused on the past, making only 3 out of 12 requests for prospectively relevant information, com-

pared with 29 out of 36 future-oriented information requests across the other three conditions. A more specific examination of the data showed that the shift from prospective to retrospective focus mainly involved the R&D report and the 5-year forecast. Under conditions of failure, high-responsibility subjects tended to substitute the R&D report for the 5-year forecast in their information requests. The hypothesis was, therefore, supported.

The information requests were further examined by conducting a MANOVA, using the same 2 × 2 design, on the pretest means of the four scales substituted for the type of information requested by each subject. For example, if a subject chose the R&D prospectus, we used the mean ratings of the R&D prospectus shown in Table 1 (i.e., 5.9, 3.8, 5.0, and 3.6) as the dependent variables for that subject in the MANOVA. This analysis, although bounded by the accuracy of the substituted means, gave a further indication of the extent to which subjects were focusing on justification of past actions versus future payoffs. The results showed a significant main effect for outcome, multivariate $F(4, 41) = 3.1, p < .05$, and a significant interaction between outcome and responsibility, multivariate $F(4, 41) = 7.20, p < .001$. The multivariate main effect was accompanied by significant univariate effects only on Items 1 and 3 shown in Table 1, and indicated that subjects in the failure conditions selected information that was more relevant to justifying the past. The emphasis on prospective concerns indicated by Items 2 and 4 in Table 1 did not vary significantly as a result of the

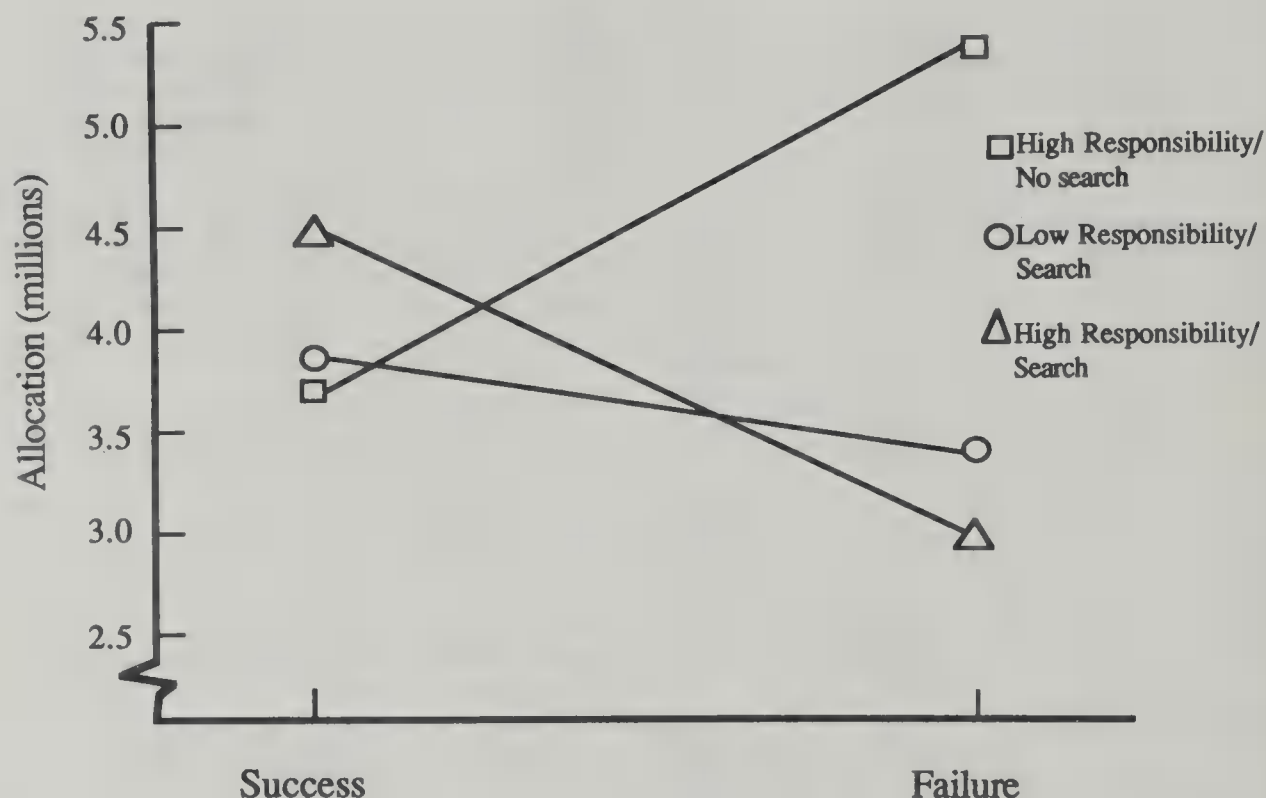


Figure 1. The effects of outcomes on allocations in various conditions of search and responsibility.

success and failure manipulation. The multivariate interaction was accompanied by significant univariate effects on all four measures and indicated that subjects who were responsible for the failure were significantly more concerned with justifying the past and less concerned about optimizing the future than were subjects in the other three conditions.

Allocations

This study also explored the effect of the search procedures on allocations. To examine this issue, we conducted an analysis of variance (ANOVA), with unique sums of squares partitionings, on the complete design. The mean allocations are listed in Table 2 and plotted in Figure 1. The results of the ANOVA are presented in Table 3. The analysis suggested that the typical escalation effect, in which high-responsibility subjects allocate more following failures than successes, was modified by the introduction of the information-search procedures. In particular, the marginally significant Search \times Outcome interaction, $F(1, 66) = 3.68, p < .06$, which was nested within the high-responsibility condition, showed that individuals responded differently to failures when search procedures were implemented than when they were not (see Figure 1). In the no-search condition, the pattern of means over the outcome conditions resembled the typical escalation result, whereas the pattern in the search condition indicated withdrawal. Most important, the difference between the allocations made in the two failure conditions was clearly significant, $F(1, 66) = 4.25, p < .05$. This result showed that, although our implementation of information-search procedures affected the allocations as anticipated, the manipulations of success and failure were sufficient to produce the escalation effect.

We also explored the relation between the information requested (i.e., prospective or retrospective) and the amount allocated across the experimental conditions. This was done by treating the 0–1 information-request measure as an independent variable in the experimental design. Using a $2 \times 2 \times 2$ ANOVA that added the request measure as a factor to the responsibility and nonnested outcome factors, we found two significant effects: a significant main effect for information requested on allocation, $F(1, 40) = 7.4, p < .01$, and an interaction effect between search and outcome, $F(1, 40) = 4.04, p < .05$, on allocations. The main effect meant that subjects who chose but did not receive retrospective information allocated less (2.20) than those who requested but did not receive prospective information (4.52). The significant interaction indicated that the difference in the amount allocated between prospective and retrospective information requestors was significantly larger in the failure condition (difference = 2.67) than in the success condition (difference = 0.60).

Discussion

Previous research on escalation has clearly demonstrated that allocators who are responsible for failure exhibit a greater tendency to continue a failing course of action than do those who are not responsible for failure. This study extended that research by examining the cognitive processes that mediate this tendency. In particular, we tested Staw's (1981) theory about retrospective focusing. Consistent with that theory, this study showed that decision makers who perceived that they were responsible for failure tended to request information that was judged by a pretest sample to be more useful for the justification of past outcomes than for the prediction of future outcomes.

Table 3
Analysis of Variance of Effects of Search

Source	df	MS	F	p
Within-cells error	66	708.54	—	—
Search with responsibility (SR)	1	705.33	0.99	.322
Responsibility (R)	1	513.78	0.73	.398
Outcome (O)	1	117.36	0.17	.685
O × SR	1	2610.75	3.68	.059
O × R	1	96.69	0.14	.713

Subjects who were not responsible for failure tended to request information that was judged in the pretest to be useful for predicting future outcomes. Using Staw's terminology, the latter appeared to be *prospectively rational*, whereas the former were *retrospectively rational*.

In addition to yielding information about the escalation process, the information-request findings also concern the acquisition and storage of information within organizations (i.e., organizational memory). The pattern of information requests observed in this study show that the type of information (e.g., prospective or retrospective) that managers may collect or create is affected by responsibility and failure. These patterns imply that the content of the employer's data archives may be affected by factors like responsibility and failure that, by normative standards, should play no part in organizational decision making. Once in the archives, such data and reports are often available at little or no cost to subsequent decision makers and, therefore, may have a biasing influence on future decisions.

In this study we also explored the relations between information search and allocations. The findings indicated that the introduction of information search significantly altered the relation between the experimental manipulations of responsibility and outcomes and the dependent variable of allocation. We permitted subjects to request information, but we did not provide them with the information they requested. Although this tactic had the effect of both preserving the ambiguity of the situation and eliminating any effect that information content may have had on allocation, it also tended to sensitize subjects to the lack of available information. We found that subjects in the high-responsibility failure condition—in which escalation usually occurs—allocated significantly less funds when information search was implemented than when it was not. Furthermore, when search was not implemented, allocations following failure in the high-responsibility condition exceeded those made following success, just as occurred in the original study (Staw, 1976). This trend was reversed when search was implemented. The pattern suggests that search, followed by nonreceipt of requested information, caused subjects to become more cautious about making further allocations when compared with those not sensitized by the search procedures. It is important to recognize that the introduction of information requests in this study did not reduce the need for subjects in the high-responsibility failure condition to justify their past decisions. The only difference between the search and no-search subjects in that condition was the requesting and subsequent denial of information in the search condition.

Although the data gathered in this study do not lead to a conclusive explanation, the relation obtained between the information requested and the subsequent allocations made in this study could be interpreted in a number of ways. One possible explanation is that subjects who were responsible for failure had a greater need to succeed on the next decision. Once sensitized to the relative lack of information available for the decision, they chose to withdraw rather than to risk exposure to an additional failure. Another possible explanation of the results is that the ability to request retrospective information provides a kind of catharsis for subjects so that they no longer feel committed to the chosen course of action. Further research would be necessary to adequately explain the relation between search and allocation.

Directions for Research

We suggest several additional research directions. Even though the present study identified a relation between information requested and allocations, the relation was probably highly determined by the consequent unavailability of that information to the subjects. It is quite likely that the commitment effect induced by the responsibility and outcome manipulations was largely destroyed by the sense of uncertainty introduced by our failure to provide information. The next logical increment in the study of information processing, as it affects escalation, would be to examine how the *content* of requested information affects allocations. The true relation between the nature of information requested and allocations may only be studied in conjunction with a theory of how the content of received information would affect allocations. The present theory of escalation, outlined in the beginning of the article, suggests the manipulation of two aspects of information content: (a) the extent to which the information content exonerates the decision maker from blame for the previous failure, and (b) the extent to which the information would support, by some legitimated standard, continued investment in the course of action. Investigators should be cautioned that predicting allocations from the content of information may be complex, especially if the allocation is part of a strategy for justification or exoneration. Before the relation between information content and allocations can be explored in the escalation context, we need to develop models and hypotheses that relate information content to the processes of exoneration and justification. These models should augment and extend the prospective-retrospective distinction developed by Staw (1981) and investigated here.

It is also reasonable to ask whether these results, obtained using students as subjects, would generalize to experienced managers. Although case studies have been cited as evidence for the presence of escalation in field contexts (Conlon & Wolf, 1983; Ross & Staw, 1986; Staw & Ross, in press), the experimental research on escalation has generally used college students as subjects. An important exception is a field study by Schoorman (1986), which showed that escalation may occur in the performance appraisals of employees to whom managers are particularly committed. Regardless of Schoorman's findings, there are two possible reasons why the experimental results such as ours may fail to generalize to experienced manag-

ers. First, individual differences in education, experience, or maturity might lead experienced managers to respond differently to failure. Second, aspects of organizational contexts, such as management incentives or monitoring of behavior, might affect how managers respond to failure. We view the results of this study as a demonstration that Staw's theory can describe the behavior of decision makers, but we make no claims about the generalizability of these results to other populations of subjects or settings. At minimum, future researchers may address the generalization of these and other escalation results by conducting experimental research using experienced managers as subjects and by manipulating theoretically important contextual factors.

References

- Agresti, A. (1984). *Analysis of ordinal categorical data*. New York: Wiley.
- Arkes, H., & Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Processes*, 35, 124-140.
- Bazerman, M. H., Beekun, R. I., & Schoorman, F. D. (1982). Performance evaluation in dynamic context: The impact of a prior commitment to the ratee. *Journal of Applied Psychology*, 67, 873-876.
- Bazerman, M. H., Giuliano, T., & Appleman, A. (1984). Escalation of commitment in individual and group decision making. *Organizational Behavior and Human Performance*, 33, 141-152.
- Brockner, J., Nathanson, S., Friend, A., Harbeck, J., Samuelson, C., Houser, R., Bazerman, M. H., & Rubin, J. Z. (1984). The role of modeling processes in the 'knee deep in the big muddy' phenomenon. *Organizational Behavior and Human Performance*, 33, 77-99.
- Brockner, J., & Rubin, J. Z. (1985). *Entrapment in escalating conflicts*. New York: Springer-Verlag.
- Brockner, J., Rubin, J. Z., & Lang, E. (1981). Face-saving and entrapment. *Journal of Experimental Social Psychology*, 17, 68-79.
- Christensen-Szalanski, J. J., & Northcraft, G. B. (1985). Patient compliance behavior: The effect of time on patients' values of treatment regimens. *Social Science of Medicine*, 21, 263-273.
- Conlon, E. J., & Wolf, G. (1980). The moderating effects of strategy, visibility and involvement on allocation behavior: An extension of Staw's escalation paradigm. *Organizational Behavior and Human Performance*, 26, 172-192.
- Conlon, E. J., & Wolf, G. (1983, August). *The architecture of a course of action: Case studies and their implications*. Paper presented at the meeting of the Academy of Management, Dallas, TX.
- Northcraft, G., & Wolf, G. (1984). Dollars, sense and sunk costs: A life cycle model of resource allocation decisions. *Academy of Management Review*, 9, 225-234.
- Ross, J., & Staw, B. M. (1986). Expo 86: An escalation prototype. *Administrative Science Quarterly*, 31, 274-297.
- Rubin, J. Z., & Brockner, J. (1975). Factors affecting entrapment in waiting situations: The Rosencrantz and Guildenstern effect. *Journal of Personality and Social Psychology*, 31, 1054-1063.
- Schoorman, F. D. (1986, August). *The unanticipated consequence of supervisor participation in hiring and promotion decisions: The escalation bias*. Paper presented at the meeting of the Academy of Management, Chicago, IL.
- Staw, B. M. (1976). Knee-deep in the big muddy: A study of escalating commitment to a chosen course of action. *Organizational Behavior and Human Performance*, 16, 27-44.
- Staw, B. M. (1981). The escalation of commitment to a course of action. *Academy of Management Review*, 6, 577-587.
- Staw, B. M., & Fox, F. (1977). Escalation: Some determinants of commitment to a previously chosen course of action. *Human Relations*, 30, 431-450.
- Staw, B. M., & Ross, J. (in press). Behavior in escalation situations: Antecedents, prototypes and solutions. In L. L. Cummings & B. M. Staw (Eds.), *Research in organizational behavior* (Vol. 9). Greenwich, CT: JAI Press.
- Taylor, S. E., & Fiske, S. T. (1981). Getting inside the head: Methodologies for process analysis. In J. Harvey, W. Ickes, & R. Kidd (Eds.), *New directions in attribution research* (Vol. 3, pp. 459-524). Hillsdale, NJ: Erlbaum.
- Thaler, R. (1980) Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization*, 1, 39-60.

Appendix

The menu choices and associated text descriptions that were rated by the pretest sample and subsequently presented to the subjects on the computer menu were as follows.

1. Five-year sales and earnings forecast. This is a statistically based forecast of the division's projected sales and earnings over the next 5 years.
2. CEO's performance reports. This report summarizes all of the feedback that the division received from the CEO about its performance.
3. Report on R&D projects (1980-1985). This report, compiled by the divisional R&D director, discusses the results of the R&D projects over the last 5 years and provides reasons for successes and failures.

4. Justification of the R&D expenditures. This is a summary of all of the memos and reports written to justify the expenditure to the division 5 years ago.

5. Future R&D prospectus. This is a report compiled by the divisional R&D director detailing a 5-year plan for R&D expenditures and their expected results.

Received August 25, 1986

Revision received February 2, 1987

Accepted December 5, 1986 ■

Effects of Gain and Loss Decision Frames on Risky Purchase Negotiations

Paul H. Schurr

School of Business Administration, University of North Carolina at Chapel Hill

Tversky and Kahneman (1981) and others have proposed that preference for risky alternatives is influenced by positive and negative decision frames. However, such effects have yet to be examined wherein groups negotiate agreements from a continuum of risky alternatives. A total of 26 teams of Master of Business Administration (MBA) students in Experiment 1, and 12 teams of professional buyers in Experiment 2, were assigned to one of two conditions in a bargaining task. Condition 1 teams, presented with payoff tables showing chances of obtaining net profits, were induced to think of potential gains (a positive frame). Condition 2 teams, presented with chances of incurring expenses, were induced to think of potential reduced losses (a negative frame). The potential profit in both conditions was held constant. Bargaining teams thinking "a gain is at stake" were hypothesized to make less risky bargaining agreements than opposing teams thinking "a loss-reduction is at stake." This hypothesis was supported. Results further support and extend findings on the causal relation between positive and negative decision frames and judgment.

In recent years, attention has focused on biases in judgment that affect decision makers (Nisbett & Ross, 1980). Although it comes as no surprise that judgment bias is also important to the understanding of negotiated agreements (Bazerman & Neale, 1982, 1983), relatively little experimental research has tested such influences in bargaining. The bargaining experiments reported here address this deficiency in bargaining research. The experiments are related to the interesting observation that individuals make different decisions depending on whether they view a choice problem as the prospect of a gain or the prospect of a loss (Bazerman, 1983; Kahneman & Tversky, 1979, 1982; Thaler, 1980; Tversky & Kahneman, 1981). Specifically, individuals appear to be more risk averse with respect to the prospect of a gain, as compared to the prospect of a loss. The objective here is to examine an application of this apparent judgment bias to a bargaining context.

Judgment bias induced by focusing a decision-maker's attention on gains or losses in a choice problem has been conceptualized in terms of a *decision frame*. The term *decision frame* refers to a "decision-maker's conception of the acts, outcomes, and contingencies associated with a particular choice" (Tversky & Kahneman, 1981, p. 453). A decision frame influences an individual's perspective on a choice problem and may determine not only the relative importance of different issues, but also the characteristics of a final decision or agreement. The frame adopted by a decision maker is influenced partly by the formulation of a problem and partly by habits and personal characteristics.

For example, Tversky and Kahneman (1981) posed a problem to individuals in which two alternative programs are proposed to combat an unusual and deadly Asian disease. One program has a certain outcome and the other has a chance outcome with an equivalent expected value. When the alternative programs are stated in terms of lives saved (a gain), 72% of the subjects prefer the certain alternative. When the alternatives are framed as lives lost (a loss), 78% prefer the risky program, even though the net outcome is identical for both decision frames (i.e., lives saved vs. lives lost). The theoretical explanation for these interesting framing effects is that individuals are inherently more averse toward placing a gain in jeopardy than they are toward risking a loss (Green, 1963; Halter & Dean, 1971; Kahneman & Tversky, 1979; Payne, Laughunn, & Crum, 1980, 1981; Swalm, 1966; Thaler, 1980, 1982; Tversky & Kahneman, 1981).

To extend the framing idea to purchase negotiations, at least two questions must be addressed. First, what frames are naturally found in business? Two frames analogous to Kahneman and Tversky's (1979) gain or loss frames are net profit and expense frames (cf. Bazerman, Magliozzi, & Neale, 1983, cited in Bazerman, 1983). Consider a situation where net profit equals gross profit less expenses. If a bargainer focuses on the net profit associated with alternative agreement levels, then his or her perspective is framed in terms of net monetary gain. Conversely, focusing on the expense associated with alternative agreement levels frames a bargainer's perspective in terms of monetary decrease or loss. Note that only the perspective, not the amount, of information differs between net profit and expense frames in this situation (i.e., profit and expense information is complementary when gross profit is held constant). Only the perspective (decision frame) differs.

Second, how can net profit and expense decision frames affect bargaining processes and outcomes? Essentially, the idea explored here is that a decision frame affects a bargainer's accep-

The author gratefully acknowledges the helpful comments of the anonymous *Journal of Applied Psychology* reviewers.

Correspondence concerning this article should be addressed to Paul H. Schurr, who is now at the School of Business Administration, State University of New York, 1400 Washington Avenue, Albany, New York 12222.

tance of risk associated with alternative levels of agreement. In theory, a net profit frame (alternatives presented as the prospect of gain) causes a bargainer to seek less risky alternatives, and an expense frame (alternatives presented as the prospect of loss) causes a bargainer to accept riskier alternatives. When two such bargainers vie for their share of a fixed-size pie (i.e., a zero-sum or constant-sum game), the bargainer with the positive (gain-anchored) frame will seek better chances for obtaining net profit, whereas the bargainer with the negative (loss-anchored) frame will yield and accept poorer chances for reducing expenses. Hence, the bargainer with the gain-anchored frame comes out ahead.

The effect of the gain or loss frames on risk preference was tested by examining bargaining outcomes as well as a bargainer's initial reservation point prior to bargaining. A reservation point is an outcome level at which a bargainer is indifferent to either making agreement or foregoing agreement (cf. Raiffa, 1982; Tedeschi, Schlenker, & Bonoma, 1973; Walton & McKersie, 1965). An initial reservation point reflects a bargainer's assessment of alternative agreement levels before bargaining begins. Thus, measurement of a bargainer's reservation point can be used to assess initial preferences for risky alternatives.

Furthermore, a reservation point indicates a bargainer's aspiration level for the outcomes of a negotiation (cf. Siegel & Fouraker, 1960). Information on aspiration levels compared with actual outcomes of negotiation permits inferences about bargainers' changing expectations during the course of negotiation.

Note that in comparison to Tversky and Kahneman's (1981) framing problems, preferences for outcomes associated with different degrees of risk, not the presence of risk aversion versus risk seeking, is of interest here. This difference stems from an adaptation of Tversky and Kahneman's (1981) framing methods to a bargaining context in which negotiators consider many different alternative levels of agreement for several different issues, as explained in the Methods section.

Reservation point hypothesis. Bargainers whose payoffs are framed in terms of net profit will prefer less risky reservation points than will bargainers whose payoffs are framed in terms of expenses.

Expected profit hypothesis. Bargainers whose payoffs are framed in terms of net profits will obtain less risky outcomes than will bargainers whose payoffs are framed in terms of expenses.

Correlation between reservation points and expected profits. Reservation points will correlate positively with outcomes attained in bargaining. Less risky reservation points will correspond to less risky outcomes.

A positive correlation between reservation points and bargaining outcomes is hypothesized because a reservation point indicates a bargainer's aspiration for the final agreement.

Method

Subjects

The experiment was run twice, once with Master of Business Administration (MBA) students and once with members of the National Association of Purchasing Managers. Details on each group of subjects are provided in the respective Results sections. The following method was used both times the experiment was run.

Design

Two decision frame treatments, expense and net profit frames, were studied across two successive trials, called *negotiating sessions*. The decision frame treatments were manipulated in a between-subjects design. Thus, the two successive trials test the reliability of the effects. Subjects bargained in teams, as will be explained. Teams in the net profit frame condition always bargained with teams in the expense frame condition. Teams never saw the frame of their opponents.

Procedures

Arrival. Subjects assembled in a central room, where they read instructions for the Commodity Broker Negotiation Game. First, they were told the objective of the game:

The objective of this game is to make good decisions about risky alternatives and to persuade your opponents to make an agreement that is to your advantage in terms of maximizing your team's profit. By maximizing your team's profit, your team will "win" the game.

Teams. Next, the written instructions indicated the role of teams:

Green Teams . . . negotiate with Buff Teams. The negotiations will actually be face-to-face. . . . The team effort is all important, since it is the team's profit that determines the winners.

Context for negotiation. Each team played the role of a committee of brokers who negotiate with another committee of brokers.

All participants in these negotiations are commodity brokers. As you probably know, a broker is an agent who negotiates contracts of purchase and sale. Although many brokers negotiate for a fee, you are a special kind of broker. You buy and sell on speculation with the expectation that you can modify the goods so that they will yield a profit. This speculative buying and selling causes a great deal of uncertainty about your future profits.

The game was designed around a broker role to avoid casting opposing bargainers as strictly buyers or sellers. A pretest showed that subjects have a tendency to adopt presumed norms for buyers or sellers. Differences associated with such norms would be confounded with the frame inductions.

As brokers, the subjects were instructed to negotiate quality levels, which are like different grades of lumber or oil. For instance, Saudi-Arabian crude oil is generally the highest grade. Furthermore, subjects negotiated quality levels for three different commodities, which correspond to three different exchanges, say, oil from Texas, Saudi Arabia, and Mexico. All of the participants had information about payoffs for different quality levels contained in tables similar to Tables 1 and 2.

You plan to negotiate ■ specific quality level for each commodity. Once your opponents have agreed to a quality level for one or more commodities, you have concluded a deal. For each commodity quality level, however, you stand the chance of making one of two levels of net profit. You might think of each transaction as involving some commodity that has to be reworked or reprocessed. At Quality Level 1, Buff Teams take the chance of experiencing low yields from reprocessing and correspondingly high costs. At Quality Level 8, Green Teams take the chance of experiencing low yields and high costs. However, the chances and costs for each side vary in a different manner from Levels 1 to 8. This means Green and Buff teams may actually agree at any Quality Level from 1 to 8, depending on the commodity. . . . You may assume that the other terms of agreement, such as price and delivery, are fixed at prevailing market rates. Your task as a broker is to obtain a favorable quality level.

Table 1
Net Profit Frame Payoff Chances for Green Teams

Quality level	Commodity A:	Commodity B:	Commodity C:
	% chance for \$1,000 net profit/unit (stake = \$1,000/unit gross profit)	% chance for \$1,000 net profit/unit (stake = \$1,000/unit gross profit)	% chance for \$100 net profit/unit (stake = \$100/unit gross profit)
1	80% chance . . . \$1,000/unit 20% chance . . . zero	70% chance . . . \$1,000/unit 30% chance . . . zero	80% chance . . . \$100/unit 20% chance . . . zero
2	70% chance . . . \$1,000/unit 30% chance . . . zero	60% chance . . . \$1,000/unit 40% chance . . . zero	70% chance . . . \$100/unit 30% chance . . . zero
3	60% chance . . . \$1,000/unit 40% chance . . . zero	50% chance . . . \$1,000/unit 50% chance . . . zero	60% chance . . . \$100/unit 40% chance . . . zero
4	50% chance . . . \$1,000/unit 50% chance . . . zero	40% chance . . . \$1,000/unit 60% chance . . . zero	50% chance . . . \$100/unit 50% chance . . . zero
5	40% chance . . . \$1,000/unit 60% chance . . . zero	30% chance . . . \$1,000/unit 70% chance . . . zero	40% chance . . . \$100/unit 60% chance . . . zero
6	30% chance . . . \$1,000/unit 70% chance . . . zero	20% chance . . . \$1,000/unit 80% chance . . . zero	30% chance . . . \$100/unit 70% chance . . . zero
7	20% chance . . . \$1,000/unit 80% chance . . . zero	10% chance . . . \$1,000/unit 90% chance . . . zero	20% chance . . . \$100/unit 80% chance . . . zero
8	10% chance . . . \$1,000/unit 90% chance . . . zero	0% chance . . . \$1,000/unit 100% chance . . . zero	10% chance . . . \$100/unit 90% chance . . . zero
100% chance alternative	\$250/unit net profit	\$250/unit net profit	\$25/unit net profit

Note. The payoff tables actually were presented to subjects in a more explicit format as follows: 80% chance net profits will equal \$1,000/unit, and 20% chance net profits will equal zero.

Like most games, the Commodity Broker Negotiation Game is an abstraction that is founded in reality. In real life, for instance, a broker might act as a speculator. The broker arranges a transaction in which a material is acquired, perhaps in exchange for another material, rather than money. Then the broker commissions the material to be reprocessed, hopefully increasing its value before resale. It is not unlikely that a higher quality material will have a higher yield, therefore leading to

more profit. Subjects who requested further clarification of their roles received explanations along these lines.

Details on how a team wins. Participants next received detailed information on how the winning teams were determined.

Because there are different schedules for green and buff teams, there will be two winning teams: one buff and one green. (Neverthe-

Table 2
Expense Frame Payoff Chances for Buff Teams

Quality level	Commodity A:	Commodity B:	Commodity C:
	% chance for \$1,000 expenses/unit (stake = \$1,000/unit gross profit)	% chance for \$1,000 expenses/unit (stake = \$1,000/unit gross profit)	% chance for \$100 expenses/unit (stake = \$100/unit gross profit)
1	90% chance . . . \$1,000/unit 10% chance . . . zero	100% chance . . . \$1,000/unit 0% chance . . . zero	90% chance . . . \$100/unit 10% chance . . . zero
2	80% chance . . . \$1,000/unit 20% chance . . . zero	90% chance . . . \$1,000/unit 10% chance . . . zero	80% chance . . . \$100/unit 20% chance . . . zero
3	70% chance . . . \$1,000/unit 30% chance . . . zero	80% chance . . . \$1,000/unit 20% chance . . . zero	70% chance . . . \$100/unit 30% chance . . . zero
4	60% chance . . . \$1,000/unit 40% chance . . . zero	70% chance . . . \$1,000/unit 30% chance . . . zero	60% chance . . . \$100/unit 40% chance . . . zero
5	50% chance . . . \$1,000/unit 50% chance . . . zero	60% chance . . . \$1,000/unit 40% chance . . . zero	50% chance . . . \$100/unit 50% chance . . . zero
6	40% chance . . . \$1,000/unit 60% chance . . . zero	50% chance . . . \$1,000/unit 50% chance . . . zero	40% chance . . . \$100/unit 60% chance . . . zero
7	30% chance . . . \$1,000/unit 70% chance . . . zero	40% chance . . . \$1,000/unit 60% chance . . . zero	30% chance . . . \$100/unit 70% chance . . . zero
8	20% chance . . . \$1,000/unit 80% chance . . . zero	30% chance . . . \$1,000/unit 70% chance . . . zero	20% chance . . . \$100/unit 80% chance . . . zero
100% chance alternative	\$750/unit expense	\$750/unit expense	\$75/unit expense

Note. The payoff tables actually were presented to subjects in a more explicit format as follows: 90% chance expenses will equal \$1,000/unit, and a 10% chance expenses will equal zero.

less, buff and green teams compete. The winning buff team, for example, will win at the expense of the two green teams it faces. Similarly, the winning green team will win at the expense of its buff opponents.) To determine your team's profit . . . the net profit from each transaction will be totaled. The green team and the buff team with the greatest total net profit for two negotiations will be designated winners.

Explaining chance outcomes. Although the payoff tables appear somewhat complicated at first (Tables 1 and 2), they are actually simple to use once participants have worked through a familiarity task. Consider Commodity A, for example. At each quality level in Table 1, the payoff is either a \$1,000 profit or no profit, depending on chance. Similarly, Table 2 gives the chance for a \$1,000 expense or no expense. No expense corresponds to \$1,000 profit (the stake is \$1,000 gross profit, less expenses). Thus, it is the chance of obtaining a profit that varies, not the amount of the payoff that varies. The lottery aspect of the chances were explained in detail in a later familiarity task (see ahead).

Interest in the task was enhanced by two variations in the commodities. Commodity B has different chances for a favorable outcome, and Commodity C has a \$100/unit stake instead of a \$1,000/unit stake.

Explaining the certain alternative. Like Tversky and Kahneman's (1981) decision problem that was presented earlier in the article, a certain alternative was provided along with the chance outcomes:

If you choose not to make an agreement with your opponent for one or more commodities, then the risk-free "Available Alternative" will apply in each instance when we calculate your team's total net profit.

In Tversky and Kahneman's decision problems, a subject must choose between either the certain or the risky alternatives. In this bargaining game, a team could choose between a risky agreement or the certain alternative, but the choice was not binary. A risky agreement—one with a chance of \$1,000 profit—presumably would be more attractive than the certain alternative if the chances of a favorable outcome appeared to be acceptable. Thus, participants in the game were motivated to closely examine the chances for getting \$1,000 profit at different quality levels.

It is important to note that a dyad's choice of the certain alternatives reduces the chance of observing differences between treatments, because when a dyad opts for the certain alternatives, a pair of expense and net profit frame teams receive precisely the same payoff, thus neutralizing the differences between treatments. This neutralizing effect is appropriate for the purposes of the study, because teams that opt for the certain alternatives are not necessarily expressing risk-taking behavior, but more likely are revealing their inability to reach agreement with the opposing team.

Give and take. To get the most profit, teams sought the best chances for no expenses. Because better chances had more value, teams were willing to give up something to get better chances of profit. In this game, the only thing they could give up would be the quality level demanded for one of the other commodities. Thus, bargaining for green teams took the form of "I'll give you a Quality Level of 6 on commodity B, if you'll give me a Quality Level of 2 on commodity A." A buff team might reply, "I'd prefer a Quality Level of 7 on B and a 6 on A." Reaching agreement was a challenge in this game because the payoffs were inversely related for the opposing teams.

Familiarity task. Once questions about game procedures were answered, participants completed a task familiarization exercise. The details of this task are important because they reveal how we caused participants to think about the meaning of risk in the game.

First, think about the chance outcomes associated with each Quality Level (in the payoff table) as a lottery. The lottery works like this. Say there is a 30% chance for outcome X and a 70% chance for outcome Y. And imagine that in a bag of 100 ping pong balls,

30 are marked "outcome X" and 70 are marked "outcome Y." If you agree to the lottery, a ball will be drawn from the bag. If the ball is marked "outcome X," you must accept the consequences of outcome X. If it is marked "outcome Y," you must accept this outcome. (In practice, at the end of this game we will determine by lottery, using random numbers, the outcome for each Quality Level. Then we will determine the net profit for each team.)

Further directions instructed participants how to determine their reservation point for a certain versus a risky alternative. Essentially, they were asked to find the point (a specific quality level) at which they were indifferent between accepting the lottery (chances) associated with that quality level and accepting the for-certain outcome associated with the available alternative for each commodity. Those versed in utility theory will recognize this as an adaptation of a fairly well-established task for determining an individual's utility function for risky decisions (Keeney & Raiffa, 1976).

Team familiarization task. Next, subjects went to an assigned team meeting room for the purpose of (a) meeting the other team members, (b) jointly determining team indifference levels for each commodity, and (c) discussing approaches to the forthcoming negotiation. Indifference points obtained from individuals and groups were used to assess group polarization effects (Myers & Lamm, 1976). Such effects were not evidenced in this study.

Bargaining tasks. Teams were ushered to their first bargaining session with a preassigned opponent. On completion of the first bargaining session, teams were given time to regroup; then they were ushered to the second bargaining session with a different team. Different teams were faced in the second session to avoid development of norms and obligations, which would make results in the second session somewhat dependent on team-specific behavior in the first session. This dependence would reduce the value of the second negotiating session as a replication of the first. Participants filled out questionnaires at the conclusion of the bargaining and then were debriefed.

Summary. The following is an overview of the experiment. (a) Individuals receive instructions on arrival in the behavioral laboratory. (b) Individuals perform the risk-assessment familiarization task (reservation point data is obtained). (c) Individuals are directed to a team meeting room. (d) Teams repeat the risk-assessment familiarization task. (e) Teams discuss bargaining strategy. (f) Teams are moved to a room where Session 1 takes place with the first opposing team (and the first outcomes—agreements for each commodity—are obtained). (g) Teams are directed to separate meeting rooms for regrouping. (h) Teams go to Session 2 and face a different opposing team (and the second outcomes are obtained). (i) Individuals are separated and complete the postbargaining questionnaire. (j) Subjects are debriefed.

Experimental Controls

Random assignment was used to assign subjects to teams, teams to conditions, and teams to opponents. The sex and number of team members was balanced between conditions. Bargainers talked directly with their opponents to reach agreements (cf. Angelmar & Stern, 1978; Rubin & Brown, 1975, p. 99). Monitors prevented collaboration between teams.

Bargaining sessions were limited to 12 min, including a 2-min warning. Time pressure arouses group members and, at low levels of arousal, facilitates task performance (Isenberg, 1981). Time pressure can cause negotiators to be less demanding and more conciliatory in bargaining situations, which facilitates agreement (Pruitt & Drews, 1969). Pretests showed the 12-min period to be sufficient for this bargaining task.

Outcome Measurement

The outcome measure is the sum of expected profit per unit for each of the three commodities, A, B, and C. Rubin and Brown (1975) sug-

Table 3
Mean Expected Profits (in Dollars) for Experiments 1 and 2

Dependent measure	Experiment 1		Experiment 2	
	Net profit group	Expense group	Net profit group	Expense group
Reservation point				
<i>M</i>	818	788	1142	953
<i>SD</i>	273	229	352	388
Session 1 outcome				
<i>M</i>	886	736	984	608
<i>SD</i>	144	120	398	245
Session 2 outcome				
<i>M</i>	882	758	964	702
<i>SD</i>	120	88	364	211

Note. $n = 13$ for each group in Experiment 1; $n = 6$ for each group in Experiment 2.

gested that researchers have traditionally used the outcomes obtained by bargainers and the symmetry with which the outcomes are divided as reliable and valid indicators of effective bargaining. Because the binary lottery outcomes did not vary (i.e., they were \$1,000 and zero for all quality levels), the expected value essentially varies according to the chances for these binary outcomes agreed to by the teams. Thus, a high-expected profit is obtained only if the chances of a \$1,000 outcome are correspondingly high. Consequently, for the unique circumstances of these experiments, expected profit reflects the degree of risk accepted by bargainers, and higher expected profits represent less risky outcomes. Recognizing that expected profit in this experiment indicates the degree of risk accepted in an agreement, it is most consistent with the bargaining literature to use expected dollar profits as the outcome measure for analysis.

Experiment 1

Subjects

In Experiment 1, subjects were MBA students, 52 men and 22 women (balanced between conditions), who volunteered for a bargaining exercise. The subjects were in their 20s and were in the first or second year of graduate study. Subjects were randomly assigned to 1 of 22 teams with 3 members, or 1 of 4 teams with 2 members, balanced between experimental conditions.

Subjects were primarily motivated by recruiting announcements suggesting that the bargaining exercise offered the benefit of learning by doing. Subjects had incentive to perform well because the teams with the highest profit (lowest expenses) were to be publicly recognized in the business school through a prominent posting of team members and scores.

Results

Reservation point hypothesis. Operationally, net profit frame teams are hypothesized to evidence less risky (higher) reservation points, as indicated by expected profit, than are expense frame bargainers. Experiment 1 data show minimal support for the first hypothesis. The means, shown in Table 3, reveal the reservation point for net profit frame bargainers was \$30 (3.8%) higher than that indicated by the expense frame bargainers, but this difference was not statistically significant, $t(25) = .30$, $p > .05$, one-tailed test.

Expected profit hypothesis. Operationally, net profit frame teams were hypothesized to obtain less risky outcomes, as measured by expected profit, than were expense frame teams. Strong support was found for this hypothesis, $F(1, 24) = 11.8$, $p \leq .01$. (Contact the author for the analysis of variance [ANOVA] table.) The means indicate that the net profit frame bargaining teams obtained higher expected profits than the expense frame bargaining teams in both negotiating sessions (Table 3). Because the expected profits reflect differences in the probability of obtaining either \$1,000 or zero, the higher expected profit for the positive frame indicates a less risky (i.e., higher probability) outcome. The net profit frame bargainers obtained 20% higher expected profits in the first negotiating session and 16% higher expected profits in the second.

There was no negotiating session effect, $F(1, 24) = .1$, $p > .05$, nor a Decision Frame \times Negotiating Session interaction, $F(1, 24) = .3$, $p > .05$. Because the no-agreement certain alternative was recorded for only 6 of 78 joint decisions (3 commodities \times 13 dyads \times 2 sessions), it proved to be a minor factor in diminishing the framing effect. Furthermore, the time limit on the bargaining task did not appear to play a role in the results. All of the bargainers reached a conclusion to their bargaining prior to the 12-min time limit, although few sessions concluded prior to the 2-min warning. Thus, whenever the certain alternative was recorded, it was not arbitrarily assigned by the experimenter as the result of an inconclusive bargaining session (cf. Kelley, 1966).

Correlation between reservation points and expected profits. Reservation points were hypothesized to correlate positively with expected profits obtained in bargaining. Only the first negotiating session was considered in the correlation analysis because expectations for the second session would be influenced by experiences in the first negotiation. The hypothesized correlation was not supported by the data. For expense frame teams there was no reservation-point-outcome association ($r = -.03$, $p > .05$). For net profit frame teams there was a negative correlation, not an expected positive correlation ($r = -.70$, $p \leq .01$).

Discussion

The unexpected negative correlation for net profit framed teams is explained, perhaps, by a ceiling effect, because these teams on average exceeded their reservation points by \$68 ($SD = \387). Yet the extent to which a team could obtain an outcome exceeding their initial reservation point was, in fact, limited to the highest outcome specified in the payoff table. Of course, an opposing team's expectations would lower the outcome ceiling even more. A negative correlation could result if an effective ceiling on outcomes limited the extent of departure from a team's initial reservation point.

Expense frame teams by comparison had outcomes \$52 ($SD = \261) less on average than their initial reservation points. This result indicates that new information obtained during bargaining caused expense frame teams to revise their reservation points downward on average. Although it could be expected that bargainers would revise their expectations throughout the bargaining process, measurement of continually developing expectations would be problematic. In retrospect, it is not surprising that some teams were overly optimistic in their initial reser-

vation points. Because there were no established norms for settlements, a conservative reservation point, one characterized by little risk, probably seemed appropriate initially. But through the course of bargaining, teams gained information about the likelihood of agreement at different quality levels. Thus, the value of agreement had to be weighed against the value of outcomes at different levels. It is particularly interesting that expense frame teams on average accepted agreements at levels below their initial reservation points, whereas profit frame teams were less inclined to do so. This contrasting behavior indicates that the expense frame teams valued the less risky initial reservation point less than they valued reaching agreement, even though the terms of agreement represented a higher level of risk.

Experiment 2

Subjects

In all, 32 professional buyers formed 12 teams in a replication of Experiment 1. Four teams had 2 members and eight teams had 3 members, balanced between gain and loss conditions. The subjects were asked to participate in the negotiation simulation as part of a week-long executive education program sponsored by the National Association of Purchasing Managers. All of the participants in the second study were professionally involved in organizational buying. Of the participants, 79% were men.

Like the management students, the professional buyers were motivated to perform well by appealing to their intrinsic interest in professional accomplishments. The reward for being on a winning team was formal recognition at a scheduled awards ceremony. A sense of professional pride made this an especially effective way to induce high involvement in the study. All aspects of the method were the same as in Experiment 1.

Results

Reservation point hypothesis. The numerical difference between the means for the two frame conditions was fairly large (\$189). Also, the means suggested that the positive frame induced preferences for less risky reservation points than did the negative frame, as hypothesized. However, as in the first experiment, the difference between the frame conditions was not statistically significant, $t(11) = .88, p > .05$, one-tailed test.

Expected profit hypothesis. As in the first experiment, the hypothesized effect of positive and negative frames was statistically significant, $F(1, 10) = 8.2, p \leq .05$. Averaging over both negotiating sessions, the purchasing manager teams in the net profit frame condition achieved an expected net profit totaling \$319 (48.7%) higher than expense frame teams, which supports the expected profit hypothesis that says the positive frame would induce bargainers to obtain less risky (higher probability) outcomes. The mean expected profits for the profit frame teams, as compared with the expense frame teams, were 62% higher in the first negotiating session and 37% higher in the second. There was no negotiating session effect, $F(1, 10) = .0, p > .05$, nor Decision Frame \times Negotiating Session interaction, $F(1, 10) = 0.1, p > .05$.

Correlation between reservation points and expected profits.

A nonsignificant positive correlation was observed between the reservation points and outcomes for net profit frame teams ($r = .17, p > .05$).

A nonsignificant negative correlation was found for expense frame teams ($r = -.35, p > .05$). According to the means in Table 3, on average, both the profit and expense frames led to outcomes below initial reservation points. As was the case in the first experiment, teams modified their expectations for preferred outcomes during the process of negotiation in the direction of accepting more risk. In summary, the Experiment 2 results for all three hypotheses replicated the results for Experiment 1. Although strong support was found for the expected profit hypothesis, the reservation point hypothesis and the hypothesized correlation between reservation points and expected profits were not supported.

Discussion

Earlier work by Tversky and Kahneman (1981; Kahneman & Tversky, 1979, 1982) suggested that important differences exist in how individuals respond to questions framed in terms of losses versus gains. The results found in the two studies reported here fully support Bazerman's (1983) assertion that the response differences identified by Tversky and Kahneman are "critical in describing negotiator behavior" (p. 212).

Consistent with the expected profit hypothesis, a positive, gain-oriented decision frame caused bargainers to prefer and obtain less risky outcomes than did their opponents, who were induced to think in terms of a negative, loss-oriented perspective. The finding of a gain and loss frame effect in both experiments was actually quite surprising considering the nature of the bargaining process. Bargainers often develop a shared expectation for settlement levels in the course of bargaining, so in face-to-face bargaining some convergence of outcomes is possible, even likely. It is interesting that the gain and loss frame effect persisted into the second negotiating session in both experiments and appeared quite stable.

Although the reservation point hypothesis did not receive statistically significant support, the means were in the hypothesized direction. Reservation points for the net profit (gain) frame, as compared to the expense (loss) frame, were on average 4% higher in Experiment 1 and 20% higher in Experiment 2, indicating the net profit frame caused preference for higher probability outcomes. It is important that these means are consistent with predictions based on gain and loss frames because the reservation point measure, as a prenegotiation measure, is the experiments' purest test of unambiguous causal relation for the individual effects of net profit and expense frames. Outcome measures in these experiments, on the other hand, reflect the mutual influence of both frames because both frames influenced the agreements reached by the dyads.

Although frames theoretically influence task conceptualization, the lack of association between reservation points and bargaining outcomes indicates a lesser role for the initial conceptualization of the bargaining task in explaining framing effects. Instead, a gain or loss frame more likely affects a bargainer's perspective throughout the bargaining process. The finding that teams on average agreed to outcomes poorer than their initial

reservation points is not particularly surprising because bargainers frequently revise their reservation points based on information acquired through bargaining (Kelley, 1966).

Given that the results of Experiment 2 replicated Experiment 1, it is interesting to consider how the subject pools differed beyond the student or practitioner classification. In particular, the professional buyers rated themselves as being *somewhat experienced* in negotiation, on the average, as compared to the self-rated *inexperienced* MBA students.

Although negotiation experience by itself is not necessarily a predictor of effective negotiating skills (Raiffa, 1982), individuals who have had business-related bargaining experience may have different risk-taking preferences in a bargaining situation. Because the decision frames in this study directly affect risk-taking behavior, replication of the framing effect among more experienced negotiators suggests the characteristics associated with experience do not negate the framing influence.

Replication of the framing effect in the purchasing manager sample is probably the most positive statement about both the internal and external validity of the studies. With respect to internal validity, experimental control was reduced by the use of groups as the decision-making unit and the presence of face-to-face communications in the course of the bargaining game. However, these factors increased external validity and did not prevent replicating the framing effect.

Comparisons to Other Research

It is important that the judgment bias caused by net profit and expense frames in a bargaining context operates in the same direction as the Kahneman and Tversky (1979, 1982; Tversky & Kahneman, 1981) gain and loss effect: Less risky decisions occur when a net profit (gain) is at stake than when a reduced expense (loss) is at stake. In fact, these findings are a significant extension of Kahneman and Tversky's work and other research inspired by their work, because the present results focus on degrees of risk taking (choice from a continuum of risky outcomes). Previous work centered exclusively on risk seeking (choice of a risky alternative) versus risk aversion (choice of a certain alternative). For this reason, the studies reported here break new ground and are not directly comparable to most previous research.

A possible exception is a study conducted by Bazerman et al. (1983; described in Bazerman, 1983). In this study the gain and loss frame was operationally defined in terms of net profits and expenses, as in the present study. The Bazerman et al. study created a simulation in which the number of transactions completed and overall profitability were examined. Because the positively framed (gain-oriented) bargainers scored higher on these two measures, the authors concluded that positively framed bargainers experienced sufficient risk aversion to cause more compromises than in the case of negatively framed bargainers. It is interesting that the two studies reported here suggest that positively framed bargainers were less willing to compromise, inasmuch as they held out for the more favorable outcomes. This apparent contradiction may be explained by considering that positively framed bargainers have higher aspiration levels for their outcomes, as suggested by the reservation points in the two studies reported here. Higher aspirations may cause a

bargainer to strive for higher goals, but also to compromise when such behavior is necessary for agreement. To resolve this issue, future research must sort out decision frame influence on behavior aimed at coordination and compromise when there is also incentive to compete (cf. Pruitt, 1981).

The attempt in these studies to correlate bargaining outcomes with reservation point measurements may be compared to the work of Eliashberg, LaTour, Rangaswamy, and Stern (1986). They reported that the successful use of group decision theory and Nash's bargaining solution to predict bargaining outcomes from measured utility functions. Their work may suggest the next step in examining how gain and loss frames affect individual and group utility functions and bargaining outcomes.

A further connection worth considering is the linkage between decision frames as schemata and other work on cognitive schemata. The common element is that decision frames and cognitive schemata both refer to the way individuals organize information in memory. The decision frames or schemata described here emphasize a perceptual anchor that organizes information in terms of loss or gain. Previous work suggests other types of schemata that influence judgment in organizations (e.g., Calder & Schurr, 1981; Schurr & Calder, 1986; Weick, 1979) and in other contexts (e.g., Abelson 1981).

Accumulating evidence suggests that decision frames are fundamental to making choices, especially in marketing contexts (e.g., Puto, Patton, & King, 1985; Thaler, 1980, 1982). The greatest difficulty in terms of application lies with understanding which decision frames are important in a particular context. Researchers must look to the decision task itself to discover other influences on decision processes. At this early stage of the framing research program, it would seem very appropriate to conduct descriptive as well as experimental research with a view toward discovering and classifying frames used by decision makers.

References

- Abelson, R. P. (1981). Psychological status of the script concept. *American Psychologist*, 36, 715-729.
- Angelmar, R., & Stern, L. W. (1978). Development of a content analytic system for analysis of bargaining communication in marketing. *Journal of Marketing Research*, 15, 93-102.
- Bazerman, M. H. (1983). Negotiator judgment. *American Behavioral Scientist*, 27, 211-228.
- Bazerman, M. H., Magliozzi, T., & Neale, M. A. (1983, August). *The acquisition of an integrative response in a competitive market*. Paper presented at the meeting of the Academy of Management, Dallas, TX.
- Bazerman, M. H., & Neale, M. A. (1982). Improving negotiation effectiveness under final offer arbitration: The role of selection and training. *Journal of Applied Psychology*, 67, 543-548.
- Bazerman, M. H., & Neale, M. A. (1983). Heuristics in negotiation: Limitations to effective dispute resolution. In M. H. Bazerman & R. J. Lewicki (Eds.), *Negotiating in organizations* (pp. 51-67). Beverly Hills, CA: Sage.
- Calder, B. J., & Schurr, P. H. (1981). Attitudinal process in organizations. In L. L. Cummings & B. M. Staw (Eds.), *Research in organizational behavior* (Vol. 3, pp. 283-302). Greenwich, CT: JAI Press.
- Eliashberg, J., LaTour, S. A., Rangaswamy, A., & Stern, L. W. (1986). Assessing the predictive accuracy of two utility-based theories in a

- marketing channel negotiation context. *Journal of Marketing Research*, 23, 101-110.
- Green, P. E. (1963). Risk attitudes and chemical investment decisions. *Chemical Engineering Progress*, 59, 35-40.
- Halter, A. N., & Dean, G. W. (1971). *Decisions under uncertainty*. Cincinnati, OH: South-Western.
- Isenberg, D. J. (1981). Some effects of time-pressure on vertical structure and decision-making accuracy in small groups. *Organizational Behavior and Human Performance*, 27, 119-134.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47, 263-291.
- Kahneman, D., & Tversky, A. (1982, January). The psychology of preferences. *Scientific American*, 162-170.
- Keeney, R. L., & Raiffa, H. (1976). *Decisions with multiple objectives*. New York: Wiley.
- Kelley, H. H. (1966). A classroom study of the dilemma in interpersonal negotiation. *Strategic interaction and conflict: Original papers and discussion*. Berkeley, CA: Institute of International Studies.
- Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychology Bulletin*, 83, 602-627.
- Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.
- Payne, J. W., Laughunn, D. J., & Crum, R. (1980). Translation of gambles and aspiration level effects in risky choice behavior. *Management Science*, 26, 1039-1060.
- Payne, J. W., Laughunn, D. J., & Crum, R. (1981). Further tests of aspiration level effects in risky choice behavior. *Management Science*, 27, 953-958.
- Pruitt, D. (1981). *Negotiation behavior*. New York: Academic Press.
- Pruitt, D., & Drews, J. (1969). The effect of time pressure, time elapsed, and the opponent's concession rate on behavior in negotiation. *Journal of Experimental Social Psychology*, 5, 43-60.
- Puto, C. P., Patton, W. E., III, & King, R. H. (1985). Risk handling strategies in industrial vendor selection decisions. *Journal of Marketing*, 49, 89-98.
- Raiffa, H. (1982). *The art and science of negotiation*. Cambridge, MA: Harvard University Press.
- Rubin, J. Z., & Brown, B. R. (1975). *The social psychology of bargaining and negotiation*. New York: Academic Press.
- Schurr, P. H., & Calder, B. J. (1986). Psychological effects of restaurant meetings on industrial buyers. *Journal of Marketing*, 50, 87-97.
- Siegel, S., & Fouraker, L. E. (1960). *Bargaining and group decision making: Experiments in bilateral monopoly*. New York: McGraw-Hill.
- Swalm, R. O. (1966, November/December). Utility theory—insights into risk taking. *Harvard Business Review*, 123-136.
- Tedeschi, J. T., Schlenker, B. R., & Bonoma, T. V. (1973). *Conflict, power and games: The experimental study of interpersonal relations*. Chicago: Aldine.
- Thaler, R. (1980). Towards a positive theory of consumer choice. *Journal of Economic Behavior and Organization*, 1, 39-60.
- Thaler, R. (1982). *Using mental accounting in a theory of purchasing behavior*. Unpublished manuscript, Graduate School of Business and Public Administration, Cornell University, Ithaca, NY.
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211, 453-458.
- Walton, R. E., & McKersie, R. B. (1965). *A behavioral theory of labor negotiations: An analysis of a social interaction system*. New York: McGraw-Hill.
- Weick, K. E. (1979). Cognitive processes in organizations. In B. M. Staw (Ed.), *Research in organizational behavior* (Vol. 1, pp. 41-74). Greenwich, CT: JAI Press.

Received October 15, 1984

Revision received December 19, 1986

Accepted December 28, 1986 ■

Arbitration and Distributive Justice: Equity or Equality?

William W. Notz and Frederick A. Starke
University of Manitoba, Winnipeg, Manitoba, Canada

Conflict over criteria used to allocate scarce resources is widespread in organizations. Two frequently used criteria (especially in labor-management disputes) are equity and equality. The conditions under which these two criteria are likely to be used by arbitrators were examined by investigating the effect of final offer (FOA) and conventional arbitration (CA) on the decisions and attitudes of 132 graduate and undergraduate students who acted as arbitrators in a series of wage and salary disputes between universities and their faculty associations. Subjects in the CA condition made arbitration awards that were most consistent with an equality, that is, split-the-difference decision rule ($p < .001$). Subjects in the FOA condition made awards that were most consistent with equity, that is, making reference to variables such as the cost of living and comparable salaries ($p < .05$). Because the external validity of the findings for the CA condition was a concern, a second experiment was conducted using 31 practicing arbitrators as subjects. The practicing arbitrators (like their student counterparts) made decisions that were most consistent with a split-the-difference decision rule. Several alternative explanations for these findings are considered, as are the implications for organizations.

It is difficult to conceive of any form of social organization without recourse to the concept of justice. Although the term *justice* has no single meaning, it is most frequently used in the context of organizational activity to refer to the distribution of resources and rewards. Justice-related dictates such as *fairness*, *rights*, *deserving*, and so forth are frequently used either to decide or to sanctify the allocation of resources and rewards.

Because *justice* has no single meaning, conflicts are likely to develop during any resource allocation process. These conflicts typically evoke multiple criteria for assessing distributive justice, or at least multiple interpretations of a single criterion. When the conflict is prolonged, an impartial third party is often engaged to assess the relative merits of conflicting claims over what is “just” and to impose a binding settlement (arbitration). Although arbitration is widely used, there is a lack of knowledge about the decision processes and criteria that are used by arbitrators as well as about arbitrator reliability in achieving distributive justice.

Distributive Justice and Arbitration

Deutsch (1975) and Walster and Walster (1975) propose that a society can survive only if it resolves the use of conflicting criteria in ways that maximize the well-being of individuals.

This research was supported by a Social Science and Humanities Research Council grant.

The authors would like to thank Max Bazerman, Jim Driscoll, Thomas Kochan, Nabil Elias, and three anonymous reviewers for their comments on earlier drafts of this article. We would also like to thank John Atwell for his assistance in the field study and his comments on the article.

Correspondence concerning this article should be addressed to William W. Notz, Faculty of Management, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2.

Deutsch hypothesizes that *equity* will be the basis for distributive outcomes in the economic sphere of social activity, whereas *equality* will be the dominant principle of distributive justice when the goal is to foster or maintain good interpersonal relations. There are, of course, other bases of distributive justice (e.g., need), but equity and equality seem to be the two major solutions to the distributive justice problem (Sampson, 1975).

There is little doubt that distributive justice is a major concern of arbitrators. Statements such as “this case involves a just and equitable evaluation . . . with the final award representing a fair determination of the entire issue” (arbitrator Benjamin Kirsch as quoted in Elkouri and Elkouri, 1960, p. 445) are commonplace in written arbitrator opinions. There is also little doubt that arbitrators generally interpret *justice* to mean *equity* (Bornstein, 1978; Elkouri & Elkouri, 1960; Mulcahy, 1976; Seitz, 1974). Arbitrators say that they weigh the inputs of each side in the dispute (e.g., in a labor-management dispute these might be productivity, ability to pay, comparable salaries, etc.) against the outcomes each side is seeking (e.g., labor’s demand or management’s offer). In an equity model, the value of the inputs (comparison variables) is of primary importance, and (at the extreme) the parties’ offers to the arbitrator have no effect on the decision.

Because arbitration involves economic conflict, claims that arbitrators use an equity criterion are consistent with Deutsch’s analysis. However, critics (Anderson, 1974; Donn, 1977; Nelson, 1975; Stevens, 1966; Swimmer, 1975) argue that arbitrators typically compromise and (at the extreme) employ pure equality by simply splitting the difference between the parties’ positions. They further argue that because disputants anticipate such equality decisions, not only is their motivation to negotiate a resolution on their own seriously reduced (the chilling effect), but they also become increasingly reliant on arbitration over time (the narcotic effect). These process effects of conventional arbitration (CA) have been judged to be so deleterious by critics

that final offer arbitration (FOA) has been proposed as an alternative.

The FOA decision structure is explicitly designed to prevent arbitrators from employing an equality decision rule. Instead, parties who fail to reach an agreement are required to submit their final offers to an arbitrator, who must choose either one or the other. The risks in such an all-or-nothing decision structure are intended to increase the parties' motivation to settle their own disputes and thereby to avoid the negative process effects of CA.

Clearly, the alleged superiority of FOA over CA depends on the extent to which CA arbitrators actually use an equality decision rule. Three reviews of numerous CA decisions (Bernstein, 1954; Holly & Hall, 1977; Wheeler, 1974) found that arbitrators frequently (but not invariably) compromised. Most of this research is *ex post facto* and the data are therefore very difficult to interpret. In addition, Farber (1981) has raised the interesting possibility that the real cause-effect sequence is the reverse of what is normally assumed: Bargainers anticipate the arbitrator's criterion of an equitable decision and then take their positions around that value.

A rigorous test of these hypotheses requires experimental control over the value of the comparison (equity) variables and the parties' final positions. Given constant values of comparison variables—such as the cost of living and comparable salaries—and holding the union's demand constant, a test of the equity/equality hypotheses could be arranged by comparing arbitrator decisions in a situation in which management had offered, say, a 5% increase in wages with arbitrator decisions in which management had offered, say, a 10% increase. If the equity hypothesis is correct, the comparison variables would be the only basis for the decision and the arbitration awards would be constant across the two conditions. On the other hand, the equality hypothesis would be supported if the different positions of management produced differences in arbitration awards.

The situation described in the preceding paragraph formed the core of the two studies reported here. This design might seem appropriate only in CA, where the arbitrator is free to compromise. However, Swimmer (1975) presented some data suggesting that the motivation to compromise is so powerful that even FOA arbitrators might try to compromise by balancing awards over time (the "flip-flop" effect). Dworkin (1977) and Feuille and Dworkin (1979) analyzed data from other jurisdictions but found little support for the hypothesis. By expanding the experimental situation described above into a *set* of sequential decisions, the equity/equality question in FOA can be rigorously examined.

Arbitrator Motives and Attitudes

The almost universally held attitude among arbitrators that their decisions should be "just" is usually articulated as a strong preference for an equity decision rule and an equally strong rejection of an equality decision rule. Because FOA is explicitly designed to prevent equality decisions, it might be expected that arbitrators would have positive attitudes toward FOA. However, what little empirical evidence exists suggests that the opposite is true; arbitrators dislike FOA (Nelson, 1975; Weitzman & Sto-

chaj, 1980; Williams, 1979). Treble (1986) provides evidence that FOA was first used over 100 years ago in several British industries, but it was discontinued because of opposition from arbitrators.

Why should arbitrators dislike a mechanism that is designed to encourage equitable decisions? There are at least two plausible explanations for this seeming contradiction. First, arbitrators may fear that an FOA decision will cause an extreme win-lose effect, which would reduce the losing party's commitment and satisfaction and therefore would be destructive of the parties' interpersonal relationship. Such a concern implies that a positive interpersonal climate is more important than long-term economic productivity. If correct, this explanation would be a severe blow to Deutsch's hypothesis, although it would retain the status of a justice motive.

The second explanation reduces to the self-interest of the arbitrator: Arbitrators may fear that the "loser" effect of an FOA decision will threaten their future retention as arbitrators. Such a self-interest motive would imply a preference for an equality decision criterion, because only by giving each party some of its demands (compromising) could an arbitrator avoid a negative perception from one of them.

On the basis of this analysis, we hypothesized that arbitrators would dislike FOA compared with CA and that they would perceive that FOA would produce an extreme win-lose effect (i.e., FOA "losers" would be significantly less satisfied than CA decision recipients). Both of these effects were expected to be greater when arbitrator self-interest was evoked. We further hypothesized that evocation of arbitrator self-interest would produce a shift toward equality in both CA decisions and, over time, in FOA decisions.

Experiment 1

Method

Subjects and Experimental Task

The subjects were 132 Master of Business Administration (MBA) and undergraduate students enrolled in organization theory courses. Because the students had just begun the program, they had not received any instruction in conflict management, nor were they given any prior to their participation in the experiment. They received bonus points toward their course grade for their involvement in the experiment. Subjects were assigned the role of arbitrator and made decisions in four separate cases involving wage disputes. All of the cases were variations on an actual conflict between a university and its faculty association that was ultimately resolved through arbitration. Each 750-word case summarized the two sides' arguments on four comparison variables commonly found in negotiations (Bornstein, 1978): increases in the cost of living, comparable salary increases, ability to pay, and productivity.

Variables and Design

The design was a $2 \times 2 \times 2 \times 4$ split-plot factorial design (Kirk, 1982, p. 531), with repeated measures on the last factor. The three between-subject factors were (a) the form of third-party intervention (FOA or CA), (b) the relation between the parties' offers and the value of the comparison variables (structured so that equity and equality decisions were equivalent, or structured so that an equality decision was greater than an equity decision), and (c) subjects' self-interest (the students' bonus grade points were either contingent or noncontingent on the accept-

ability of their arbitration awards). The manipulation of these between-subject factors is described in detail in the Experimental Procedure section below. The within-subject repeated factor was the case describing the wage dispute.

The first dependent variable—the arbitration decision—was measured by observing the subject's decision (a check mark for one of the parties' final offers in the FOA condition or the percentage award in the CA condition). The second dependent variable—time required to make the decision—was measured by observing the starting and ending time for each case on which the subject worked. Several other dependent variables were measured using responses from a postarbitration questionnaire: (a) the satisfaction and anxiety levels of the subjects, (b) how satisfied the subjects felt the university and faculty association would be with their arbitration decision, (c) how important it was to the subjects that the disputing parties were satisfied with their decision, and (d) the relative importance of the four comparison variables in the subject's decision. All of the dependent variables were measured for each of the four cases.

Each subject was randomly assigned to one of the eight experimental conditions and was tested individually in a nonclassroom setting. Because the number of subjects in each condition (16 or 17) was less than the number of possible combinations of case orders (24), in all conditions the cases were assigned in such a way that each appeared as equally possible in each position in the sequence (first, second, third, and fourth).

Experimental Procedure

The repeated-measures design required each subject to participate in the arbitrator role using the same decision structure (FOA or CA), the same bonus-point condition (contingent or noncontingent), and the same comparison-variables condition (equity and equality, equivalent or not equivalent) for all four cases that they arbitrated. When subjects made their decision in the first case, they completed a questionnaire that measured the dependent variables and manipulation checks. Subjects were then given the second case, and the cycle was repeated until all four cases were completed.

Form of arbitration. Subjects in the FOA condition were given written instructions indicating that they would be required to choose either the university's offer or the faculty association's offer, the one they felt was most reasonable. Subjects in the CA condition were given written instructions indicating that they were free to impose any settlement they wished, as long as they felt it was reasonable.

Equity versus equality. Equity was operationalized by providing subjects with quantitative data on two comparison variables: *increases in the cost of living and comparable salaries* (see Table 1). The data in Table 1 were presented on a case-by-case basis. Because the data were embedded in a 750-word narrative, subjects saw data for only one case, made their decision, completed the postarbitration questionnaire, and then repeated the process for the second, third, and fourth cases.

As in the real case, each side's interpretation of these quantitative variables was also given. The mean of these two positions defined the value of each variable and the overall mean of the two variables defined an "equitable" decision. (Of course, only the raw data were presented to the subjects). Qualitative arguments were also summarized for two other factors—*productivity and ability to pay*—in order to increase the realism of the case.

In the "Equality = Equity" condition (EQL = EQT), a decision to split the difference between the parties' final positions (equality) would also produce an equitable decision, because it would have the same value as the mean of the comparison variables. For example, in the Central University case, the faculty association was asking for an increase of 14.1% and the university was offering a 4.9% increase. The mean of the comparison variables in this case was 9.5%; so if the subject simply

split the difference and imposed a 9.5% award, an equitable decision would have been reached *coincidentally*.

In the "Equality > Equity" condition (EQL > EQT), everything remained as before, except that management's offer was increased. Thus, a decision to split the difference in this condition would produce an award that exceeded the mean of the comparison variables. All data from the example cited above were unchanged except that management's offer was raised from 4.9% to 8.9% (see Table 1). Simply splitting the difference here (an award of 11.5%) would have resulted in an inequitable settlement, that is, one that could not be justified by the mean value (9.5%) of the comparison variables.

Subjects in the EQL = EQT condition saw only the data indicated on that line of Table 1 for the four cases (in narrative form), whereas subjects in the EQL > EQT condition saw only the data indicated on that line of Table 1 for the four cases (again only in narrative form).

High versus low self-interest. Subjects in the low self-interest condition were told that they were to make arbitration decisions that were as reasonable as possible, but that their bonus points would not depend on their decisions. Subjects in the high self-interest condition were told that their decisions, as a set, would be analyzed by a three-person panel—composed of a government expert on labor relations, a university representative, and a faculty association representative—that would determine their bonus points. (There actually was no panel.) At the end of the experiment, subjects were thoroughly debriefed. All subjects received an equal number of bonus points (more than they thought they would receive).

Results

A preliminary analysis indicated that there were no order effects (i.e., whether a given case was first, second, third, or fourth in the sequence). Similarly, there were no interaction effects between the university (case) variable and any of the between-subject factors. Thus, these variables were ignored in subsequent analyses.

Manipulation Checks

In order to determine whether the self-interest manipulation between bonus points and decision quality was successful, subject responses to two questions were analyzed. Subjects were asked to indicate on 5-point scales both what relation they perceived between the quality of their arbitration decision and the assignment of bonus points, and how likely they believed it was that they would receive bonus points. Subjects in the high self-interest condition saw a much stronger relation between the quality of their decisions and the bonus points they would receive than did subjects in the low self-interest condition, $F(1, 124) = 19.54, p < .001$. (All p -values are two-tailed unless otherwise indicated; all analyses of variance in Experiment 1 are repeated measures.) Similarly, subjects in the high self-interest condition believed that they were less likely to receive all of the potential bonus points, $F(1, 124) = 25.26, p < .001$. As a final check, the amount of time it took subjects to make their decisions was measured. Subjects in the high self-interest condition required an average of 10.2 min to reach a decision on each case, compared with 8.13 min for the low self-interest subjects, $F(1, 124) = 14.64, p < .001$.

As a check on the EQL = EQT versus EQL > EQT manipulation, subjects were asked to indicate (on a 5-point scale) which of the parties' positions was the most reasonable. Subjects in

Table 1
Summary of Quantitative Data in Each Case

Case	Union demand	Management offer	Comparison variable				Average of union demand and management offer	Average of comparison variables
			Cost of living		Comparable salaries			
			Union interpretation	Management interpretation	Union interpretation	Management interpretation		
Northland University								
EQL = EQT	12.59	4.38	9.91	6.88	11.70	5.45	8.49	8.49
EQL > EQT	12.59	7.95	9.91	6.88	11.70	5.45	10.27	8.49
Wheatland University								
EQL = EQT	13.33	4.64	10.49	7.28	12.38	5.77	8.99	8.99
EQL > EQT	13.33	8.42	10.49	7.28	12.38	5.77	10.88	8.99
Central University								
EQL = EQT	14.10	4.90	11.12	7.71	13.10	6.10	9.50	9.50
EQL > EQT	14.10	8.90	11.12	7.71	13.10	6.10	11.50	9.50
Markham University								
EQL = EQT	14.81	5.15	11.66	8.09	13.76	6.41	9.98	9.98
EQL > EQT	14.81	9.35	11.66	8.09	13.76	6.41	12.08	9.98

Note. EQL means equality; EQT means equity. Figures are percentages.

the EQL > EQT condition (in which the university's offer was much closer to the mean of the comparison variables than the faculty association's offer was) felt the university's offer was more reasonable than did subjects in the EQL = EQT condition, $F(1, 124) = 10.19, p < .05$.

Arbitrator Decisions

The determination of the extent to which arbitrators used equity or equality as the criterion for their decisions and the conditions that might enhance or inhibit these respective dispositions was the most important thrust of this research. The most straightforward data on these questions were generated by comparing subject decisions in the CA EQL = EQT condition with the CA EQL > EQT condition. Support for the equity criterion would be evident if the same distribution of arbitrator decisions existed across these two conditions, because the overall mean of the comparison variables in the four cases (9.24%) was identical in both conditions. Support for the equality criterion would be evident if there was a positive shift in the distribution of awards from the EQL = EQT to the EQL > EQT condition. This would reflect the increased value of the university's offer (the overall mean of the university and faculty association offers for the four cases was 9.24% in the EQL = EQT condition and 11.18% in the EQL > EQT condition).

The average actual award of subjects in the EQL = EQT condition was 9.46% as compared with 10.87% for subjects in the EQL > EQT condition, $F(1, 62) = 50.14, p < .001, \omega^2 = .42$. In other words, despite the constant value of the equity criterion across the two conditions, the increase of 1.94 percentage points (9.24 to 11.18) in the mean of the parties' offers produced a corresponding increase of 1.41 percentage points in the mean of the subjects' arbitration awards. This effect supports the equality hypothesis.

As indicated earlier, the existence of equity/equality effects in the FOA condition would be expected to take a somewhat different form. One way in which equality would be manifested

would be in arbitrator compromise over time (the flip-flop effect). Because the subjects made four arbitration decisions, a flip-flop effect would mean that subjects split their choices evenly between the university and the faculty association. Although the modal response of subjects was, in fact, to divide their awards evenly (see Total column in Table 2), there were no significant differences between this and the other categories of subject responses, $\chi^2(4, N = 66) = 4.30, ns$ (data taken from student decision forms). An additional analysis comparing subjects who believed their rewards were contingent with those whose rewards were not contingent also revealed no differences, $\chi^2(4, N = 66) = 1.34, ns$.

Because the increase in the university's offer in the EQL > EQT condition brought it correspondingly closer to the mean of the comparison variables while the faculty association's offer

Table 2
FOA Decision Patterns in the EQL = EQT and EQL > EQT Conditions

Decision pattern ^a	No. of subjects with each decision pattern		
	EQL = EQT	EQL > EQT	Total
Union offer chosen 4 times	9	2	11
Union offer chosen 3 times, university one time	6	3	9
Union offer chosen 2 times, university 2 times	9	10	19
Union offer chosen 1 time, university 3 times	4	9	13
University offer chosen 4 times	5	9	14
Total	33	33	66

Note. FOA = final offer arbitration. EQL means equality; EQT means equity.
^a Each subject made four arbitration decisions.

Table 3
Attitudinal Data on CA and FOA

Attitudinal data	<i>M</i>	
	CA	FOA
Effects on arbitrators ^a		
Satisfaction with procedure	3.43	2.89
Comfortable with procedure	3.19	2.66
Decision difficulty	2.73	2.63
Possibility of equitable decision	3.28	2.53
Perceived effect on disputants ^b		
Commitment index	9.31	13.90 (winners) 5.70 (losers)
Use of decision criteria ^c		
Cost of living	44.4	32.7
Comparable salaries	29.0	28.0
Productivity	6.0	10.0
Ability to pay	16.9	25.1
Other	3.7	4.2

Note. CA = conventional arbitration; FOA = final offer arbitration.
^a Scale range = 1–5. ^b based on 3 scales, each scale range = 1–5 (composite scale range = 3–15). ^c relative weights assigned out of 100%.

was left unchanged, an equity effect would imply a shift in the distribution of awards in favor of the university. Alternatively, a finding of no difference between the two conditions would support the equality hypothesis. As shown in Table 2, FOA subjects in the EQL > EQT condition selected the university's offer significantly more often than subjects in the EQL = EQT condition, $\chi^2(4, N = 66) = 8.57, p < .05$, one-tailed. These data contrast sharply with those from the CA condition where the awards shifted toward the faculty association's position. It had also been expected that the effects of the two forms of arbitration would vary with the subjects' self-interest; however, the self-interest variable had no impact on either the FOA or CA decisions.

Arbitrator Motives and Attitudes

The analyses of arbitrator motives and attitudes focused on the differences that were expected to be evoked by the FOA and CA procedures. Three areas were examined.

The effect on arbitrators. Compared with CA arbitrators, FOA arbitrators (a) were less satisfied with the procedure they used, $F(1, 124) = 12.74, p < .001, \omega^2 = .08$; (b) were less comfortable with the procedure they used, $F(1, 124) = 13.58, p < .001, \omega^2 = .09$; and (c) felt that CA would have permitted them to make more equitable decisions, $F(1, 124) = 44.35, p < .001, \omega^2 = .25$ (see Table 3 for means). Cronbach's alpha values for subjects' attitudinal and self-report data ranged from .65 to .92 ($M = .84$).

The perceived effect on the disputants. The hypothesized win-lose effect was tested by examining subject predictions of commitment both for FOA winners and conventionals, and for FOA losers and conventionals. Three highly correlated attitudinal responses were combined into an equally weighted index of predicted commitment (with a scale range of 3 to 15). Subjects predicted that FOA winners would be more committed than conventionals, and that FOA losers would be less committed

than conventionals, $F(1, 526) = 1025.13, p < .001$, and $F(1, 526) = 427.20, p < .001$, respectively. Thus, the hypothesis was supported (see Table 3 for means).

The use of decision criteria. Subjects were asked to indicate (by allocating portions of 100%) the relative importance of five variables (cost of living, comparable salaries, productivity, ability to pay, and other factors) in their decisions. Subjects in the CA condition said they were influenced more by the cost of living variable than their counterparts in the FOA condition, $F(1, 124) = 17.66, p < .001, \omega^2 = .01$. Compared with CA subjects, FOA subjects claimed that productivity and ability to pay influenced their decisions more, $F(1, 124) = 5.22, p < .05, \omega^2 = .03$, and $F(1, 124) = 10.64, p < .001, \omega^2 = .07$, respectively (see Table 3 for means).

The External Validity of Experiment 1 Results

Laboratory studies inevitably gain internal validity at the expense of external validity (Cook & Campbell, 1976). In this study, the substantial difference between offers made possible an internally valid and sensitive juxtaposition of the equity and equality hypotheses, but it also weakened the external validity of the results; this threat applies particularly to the FOA findings, because large differences between offers might be expected to occur infrequently among actual bargainers. There were other threats to external validity, but generalization from student decision behavior to the decisions of practicing arbitrators seemed to be the most significant one. It would, for example, be very easy to generate plausible arguments for a substantial effect of experience on arbitrator decisions. This threat seemed crucial, and the only satisfactory way of dealing with it was to conduct a second experiment.

Experiment 2

The purpose of Experiment 2 was to replicate in part the findings from Experiment 1, using practicing arbitrators as subjects. The finding from the first study that seemed most vulnerable to an experience threat was the equity/equality test for the decisions of CA subjects. Accordingly, the second experiment replicated the EQL = EQT and EQL > EQT conditions under CA.

Method

Sixty-eight practicing interest arbitrators were sent a cover letter requesting their participation in the study, instructions for making the arbitration decision for the Central University case, and either the EQL = EQT or EQL > EQT condition (randomly assigned). The 31 arbitrators who responded had an average of 13 years of experience and had arbitrated an average of 17 cases. The experience of the practicing arbitrators is understated, because the highest category of experience (over 25 cases) was checked by more than one-third of the sample.

Results

The results from the second experiment are shown in Table 4. The practicing arbitrators' decisions for the EQL = EQT and EQL > EQT conditions have been juxtaposed with the comparable data for the student arbitrators. The means of the compar-

Table 4
Comparison of Student and Arbitrator Awards

Condition	Student awards	Arbitrator awards	Comparison variables	Faculty and university offers to arbitrator
EQL = EQT	9.46	9.76	9.50	9.50
EQL > EQT	10.87	11.89	9.50	11.50

Note. Figures are mean percentages. EQL means equality; EQT means equity.

ison variables and parties' offers for the Central University case are also shown. Overall, the results of the two experiments were strikingly similar. The difference between the means of the EQL = EQT and EQL > EQT conditions among professional arbitrators was significant, $F(1, 26) = 11.67, p < .001, \omega^2 = .31$ (one-tailed). The increase in the split-the-difference value of the two parties' offers (two percentage points from the EQL = EQT to the EQL > EQT condition) produced a corresponding shift in the arbitrators' awards of 2.13 percentage points and 1.41 percentage points in the student decisions. There was no difference between the means or the variances of arbitrator and student decisions in the EQL = EQT or the EQL > EQT condition.

There was sufficient variance within the arbitrator sample to allow some analysis of the relation between arbitrator experience and arbitrator decisions. The sample was first split into two categories: Those who had arbitrated more than 25 cases ($n = 11$) and those who had arbitrated less than 25 cases ($n = 11$). Nine arbitrators did not indicate how many cases they had arbitrated. The salary increase that could be supported by the mean of the comparison variables, in this case 9.5%, was then subtracted from each arbitrator's decision. These "deviations from equity" scores were then summed for each group and a mean score calculated. An analysis of the means for the two groups (1.40 vs. 1.63 for the experienced and inexperienced groups, respectively) revealed no significant differences.

An additional analysis of the effect of experience on arbitrator decisions was conducted by correlating the experience of arbitrators in the EQL > EQT condition with their deviations-from-equity scores. Experienced arbitrators tended to impose higher settlements than inexperienced arbitrators, but the correlation was not statistically significant. These two analyses converge and strongly suggest that arbitrator experience did not influence their decisions.

Discussion

Do arbitrators use equity or equality as their dominant decision criterion? The answer seems to depend on the form of arbitration used. Subjects who were free to fashion awards (CA) made decisions that could best be described with an equality model. In both experiments, the comparison variables that were integral to an equity decision had no substantive effect on the awards of CA arbitrators, nor was there any evidence that this lack of effect varied with subject experience. Moreover, these data cannot be explained by a reversal in the cause-effect sequence as suggested by Farber (1981).

In sharp contrast to the equality effect among CA subjects, no similar evidence could be found for FOA subjects in Experiment 1. Instead, an increase in the value of the university's offer (so that it was closer to the value of the quantitative comparison variables than was the union's offer) caused a significant increase in the frequency with which the FOAs chose the university's offer. Thus, the distribution of awards clearly reflected the influence of an equity criterion when subjects were constrained to a choice of one or the other of the offers. Furthermore, there was no evidence of any attempt by the FOAs to achieve equality over time (flip-flopping).

Although these results strongly support critics who contend that CA arbitrators compromise when making decisions, there was little support for the hypothesis that they do so out of their own self-interest. Despite the obvious success of the self-interest (reward contingency) manipulation, neither the CA nor the FOA subjects' decisions were affected. Thus, self-interest (at least over the manipulated range) was not a plausible explanation for either the observed equality effects in CA or the lack of such effects in FOA decisions.

Another motivationally oriented explanation is that CAs imposed equality awards out of a desire to create a positive interpersonal relationship between the disputing parties. The subjects' predictions of how the parties would feel about their awards could be interpreted as support for this hypothesis, because the CA subjects' equality decisions were accompanied by predictions of relatively equally satisfied parties, whereas the FOA subjects' equity decisions were accompanied by predictions of a win-lose effect. On the other hand, this explanation is inconsistent with both arbitrator claims about their decision processes and the experimental subjects' introspections in that regard.

Although these motivational explanations cannot readily account for the results, a cognitive interpretation might be plausible. Research on human inference and judgmental heuristics indicates that there are many instances in which people are unaware of their mental processes and the factors that influence their judgments (e.g., Hogarth, 1980; Kahneman & Tversky, 1973; Nisbett & Ross, 1980; Slovic & Lichtenstein, 1971). If the results of this study represent such an instance, it means that CA subjects were influenced by the parties' offers (i.e., equality) without being aware of it and that FOA subjects were similarly influenced by the quantitative comparison variables (i.e., equity) without being aware of it. Such an explanation would require that there be no relation between the amount that a CA decision differed from a split-the-difference (equality) value and the retrospective importance that was attributed to an equity criterion in the decision process. This was, in fact, the case.

This cognitive interpretation might well be a plausible explanation of phenomena that extend beyond the bounds of this particular study. Like the arbitrator subjects in this experiment, real world arbitrators also claim that their distributive decisions should and do reflect an equity model. Moreover, these lay theories of distributive justice are entirely consistent with expert theories: Decisions in the economic sphere of activity will be based on equity (Deutsch, 1975). If this theory of distributive justice is widely shared, then it should also form the basis of the expectations of most bargainers. Unfortunately, these results suggest that the expectations of CA bargainers will be violated

because CA arbitrators produced equality, not equity decisions. FOA arbitrators, on the other hand, made decisions that were consistent with equity.

In addition to these *outcome* effects of CA and FOA, it seems likely that the two decision structures may also generate *processes* that differentially affect acceptance. Bargainers who anticipate a split-the-difference (equality) arbitration decision have an incentive to exaggerate their claims in order to offset the effects of such a decision. Conversely, the FOA structure creates uncertainty and thus induces concession and compromise among bargainers. Concession and compromise require bargainers to make choices, and the committing effects of such choices are well known (Kiesler, 1971; Salancik, 1977). The commitment generated by this process is an important factor in the determination of the negotiators' acceptance of, and feelings of responsibility for, the outcomes of the conflict (Notz & Starke, 1978; Starke & Notz, 1981). Thus, FOA induces processes that have positive effects on commitment and acceptance, whereas CA tends to induce processes that are detrimental to commitment and acceptance.

An important, but frequently overlooked, implication of this line of reasoning is that distributive justice depends on more than mere adherence to a principle of benefit distribution, whether that principle is equity, equality, or something else. As in other arenas of organizational activity, the process by which allocation decisions are made will have an important impact on participant perceptions of justice (Tyler & Caine, 1981). Thus, agreements that are voluntarily arrived at are generally seen as preferable to those imposed by a third party, even though the distributed outcomes might be identical (a voluntarily arrived-at decision to split the difference, for example, would be superior to an equivalent decision imposed by a third party). The FOA and CA decision structures therefore can be expected to differentially affect distributive justice not only in terms of equity and equality, but also through the variance in the processes of bargaining and negotiation that they evoke.

References

- Anderson, A. (1974, April). *Lessons from interest arbitration in the public sector*. Address to the meeting of the National Academy of Arbitrators, Kansas City, MO.
- Bernstein, I. (1954). *The arbitration of wages*. Berkeley: University of California Press.
- Bornstein, T. (1978). Interest arbitration in public employment: An arbitrator views the process. *Labor Law Journal*, 29, 77-86.
- Cook, T., & Campbell, D. (1976). The design and conduct of quasi-experiments and true experiments in field settings. In M. Dunnette (Ed.), *Handbook of Industrial and Organizational Psychology* (pp. 223-326). Chicago: Rand McNally.
- Deutsch, M. (1975). Equity, equality, and need: What determines which value will be used as the basis of distributive justice? *Journal of Social Issues*, 31, 137-149.
- Donn, C. B. (1977). Games final offer arbitrators might play. *Industrial Relations*, 16, 306-314.
- Dworkin, J. B. (1977). Final position arbitration and intertemporal compromise. *Relations Industrielles*, 32, 250-260.
- Elkouri, F., & Elkouri, E. (1960). *How arbitration works*. Washington, DC: Bureau of National Affairs.
- Farber, H. (1981). Splitting the difference in interest arbitration. *Industrial and Labor Relations Review*, 14, 311-317.
- Feuille, P., & Dworkin, J. (1979). Does Wisconsin's final offer arbitration offer only intertemporal compromise? *Monthly Labor Review*, 102, 39-40.
- Hogarth, R. M. (1980). *Judgement and choice*. New York: Wiley.
- Holly, J. F., & Hall, G. A. (1977). Dispelling the myths of wage arbitration. *The Labor Law Journal*, 28, 344-354.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237-251.
- Kiesler, C. A. (1971). *The psychology of commitment: Linking behavior to belief*. New York: Academic Press.
- Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences*. Monterey, CA: Brooks/Cole.
- Mulcahy, C. C. (1976). Ability to pay: The public employee dilemma. *The Arbitration Journal*, 31, 90-96.
- Nelson, N. (1975). Final offer arbitration: Some problems. *The Arbitration Journal*, 30, 50-58.
- Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.
- Notz, W. W., & Starke, F. A. (1978). Final offer vs. conventional arbitration as means of conflict management. *Administrative Science Quarterly*, 23, 189-203.
- Salancik, G. R. (1977). Commitment and the control of organizational behavior and belief. In B. M. Staw & G. R. Salancik (Eds.), *New directions in organizational behavior* (pp. 1-54). Chicago: St. Clair Press.
- Sampson, C. E. (1975). On justice as equality. *Journal of Social Issues*, 31, 45-64.
- Seitz, P. (1974). Footnotes to baseball salary arbitration. *The Arbitration Journal*, 29, 98-103.
- Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, 6, 649-744.
- Starke, F. A., & Notz, W. W. (1981). Pre- and post-intervention effects of conventional vs. final offer arbitration. *Academy of Management Journal*, 24, 832-850.
- Stevens, C. (1966). Is compulsory arbitration compatible with bargaining? *Industrial Relations*, 5, 38-52.
- Swimmer, G. (1975). Final position arbitration and intertemporal compromise: The University of Alberta compromise. *Relations Industrielles*, 30, 533-536.
- Treble, J. G. (1986). How new is final-offer arbitration? *Industrial Relations*, 25, 92-94.
- Tyler, T. R., & Caine, A. (1981). The influence of outcomes and procedures on satisfaction with formal leaders. *Journal of Personality and Social Psychology*, 41, 642-655.
- Walster, E., & Walster, W. (1975). Equity and social justice. *Journal of Social Issues*, 31, 21-43.
- Weitzman, J., & Stochaj, J. (1980). Attitudes of arbitrators toward final offer arbitration in New Jersey. *The Arbitration Journal*, 35, 25-34.
- Wheeler, H. N. (1974). Is compromise the rule in firefighter arbitration. *The Arbitration Journal*, 29, 176-184.
- Williams, B. (1979). *Arbitration report on negotiations between the University of Manitoba and the University of Manitoba Faculty Association*. Winnipeg, Manitoba, Canada: University of Manitoba.

Received July 18, 1986

Revision received December 13, 1986

Accepted December 28, 1986 ■

How Important Are Dispositional Factors as Determinants of Job Satisfaction? Implications for Job Design and Other Personnel Programs

Barry Gerhart

Center for Advanced Human Resource Studies, New York State School of Industrial and Labor Relations, Cornell University

According to recent research, stable dispositional factors may result in considerable consistency in attitudes such as job satisfaction across time and situations. If true, this finding may have important implications. For example, Staw and Ross (1985) argued that "many situational changes such as job redesign . . . may not affect individuals as intended." Such personnel programs "may be prone to failure because they must contend with attitudinal consistency" (p. 478). The present article has two purposes. First, methodological and conceptual problems with the Staw and Ross assessment of the impact of situational and dispositional factors on job satisfaction are discussed. Second, given Staw and Ross's focus on job redesign, this article examines the impact on job satisfaction of changes in two very different measures of job complexity. Findings indicate that changes in situational factors such as job complexity are important predictors of job satisfaction, consistent with Hackman and Oldham's (1975, 1976) job design model. In contrast, measurement problems preclude accurate assessment of the predictive power of dispositional factors. Contrary to the concern raised by Staw and Ross (1985) and Staw, Bell, and Clausen (1986), it does not appear likely that the success of personnel programs will be significantly constrained by the influence of attitudinal consistency.

Locke (1969) defined job satisfaction as "a function of the perceived relationship between what one wants from one's job and what one perceives it as offering". Presumably, this definition points to the importance of both dispositional and situational factors as determinants of job satisfaction. In practice, however, Mitchell (1979) suggested that personality variables have received relatively little attention in empirical research on determinants of job attitudes. Similarly, Weiss and Adler (1984) argued that "researchers have barely scratched the surface on the ways in which personality constructs may enter into theoretical systems" (p. 43). There is disagreement, however, concerning the need for future research on personality or dispositional variables as determinants of job attitudes. In Mitchell's view, the "secondary role" played by dispositional variables "seems justified and necessary" (p. 247). In contrast, Weiss and Adler argued that "It is simply premature and unproductive to make any normative statements about restricting the role of personality in organizational research" (p. 2).

The role of dispositional factors or traits¹ as determinants of job satisfaction has been examined in three recent empirical studies. Pulakos and Schmitt (1983) reported that high school students' instrumentalities for job-related outcomes measured prior to taking a job were predictive of subsequent job satisfaction. In their view, this finding suggests that "personnel selec-

tion might benefit from more attention to selecting individuals who have a higher probability of being satisfied" (p. 311). An alternative interpretation, however, is that people who expected to receive relatively good jobs (i.e., had higher instrumentalities for valued outcomes) were, in fact, more likely to receive good jobs. These persons tended to be more satisfied because they received better jobs, not because of any propensity to be satisfied. To examine the impact of traits on job satisfaction, actual job outcomes must be controlled.

Staw and his colleagues conducted two studies designed to assess the impact of traits on job satisfaction. Staw, Bell, and Clausen (1986) found that adolescent "affective disposition" was correlated with adult job affect. Although Staw et al. interpreted this relation as evidence for the impact of traits on job satisfaction, the magnitude of the correlations was moderate, the sample sizes small, and control for job attributes limited, perhaps indicating a need to view these findings as suggestive only.

In contrast to the Staw et al. (1986) study, Staw and Ross (1985) sought to investigate both traits and job factors as determinants of job satisfaction, consistent with Locke's definition. They examined the temporal stability of a single, global job-satisfaction item as a function of pay change, occupational status change, and previous global job satisfaction (measured 5 years earlier). Data were from the mature men cohort of the National Longitudinal Surveys of Labor Market Experience (NLS). Staw and Ross found that "satisfaction in 1966 was the strongest and most significant predictor of 1971 job attitudes.

I am grateful to John Boudreau, Peter Dowling, Lee Dyer, George Milkovich, and Sara Rynes for helpful comments on an earlier draft of this article.

Correspondence concerning this article should be addressed to Barry Gerhart, New York State School of Industrial and Labor Relations, Cornell University, Ithaca, New York 14851-0952.

¹ The term *trait* is used as a synonym for *dispositional factors* in this article.

Neither changes in pay nor changes in job status accounted for nearly as much variance as prior job attitude" (p. 475).

Staw and Ross (1985) drew important implications from their study. They argued that "it is difficult to conclude from the present data that situational effects will supersede attitudinal consistency in most contexts" (p. 477). (Note that Staw and Ross attributed the predictive power of previous job satisfaction entirely to trait stability—the accuracy of this assumption is discussed later.) Regarding practical implications, they concluded that "many situational changes such as job redesign and organizational development may not affect individuals as they are intended" (p. 478; see Staw et al., 1986, for a similar argument).

These conclusions contrast sharply with the way many researchers view the relative importance of situational and trait factors as determinants of job satisfaction (cf. Mitchell, 1979). At an applied level, the Staw and Ross (1985) conclusions, if valid, suggest the need for a major reexamination of the value of personnel programs designed to affect worker attitudes and consequent behaviors through changes in the work environment. If Staw and Ross are correct, many such programs may be doomed to failure because worker attitudes are, to an important degree, a function of stable individual traits, not situational characteristics.

Several aspects of the Staw and Ross (1985) study, however, limit the validity of their conclusions. For example, Staw and Ross used a sample of men between the ages of 45 and 59 in 1966 and between the ages of 50 and 64 in 1971. Relative to younger workers, this older cohort is less likely to experience significant change in the work situation. In fact, using the same data set as Staw and Ross, the present author found a test-retest correlation of .84 for both pay and occupational status between 1966 and 1971. Given this high stability of situational factors, the test-retest correlation of .29 found by Staw and Ross for job satisfaction measured in 1966 and 1971 does not provide very convincing evidence of trait stability across time.

Moreover, given the high positive correlation ($r = .84$) between the component parts of both pay and occupational status in their sample, the change scores used for pay and status by Staw and Ross (1985) in their satisfaction equations may have been very unreliable. Using a formula given by Guilford (1954, p. 394), the reliability of the difference scores used by Staw and Ross can be estimated under different assumptions. If in the cross section, the reliabilities of the pay and status measures had been .95, the reliability of the difference scores would have been .69. However, .95 is probably an optimistic estimate (Jencks, 1979, pp. 328–329). The true parameter may be considerably lower (Duncan & Hill, 1985). Given the latter research, a cross-sectional reliability of .85 to .90 may be more plausible, resulting in reliabilities for the Staw and Ross difference scores ranging from .06 to .38.

Therefore, even if the measures of the pay and status components were reliable, the pay and status change scores used by Staw and Ross were probably very unreliable, thus resulting in a serious underestimation of the effects of these situational factors on job satisfaction. Under such circumstances, comparing the relative effects of situational factors and traits is not quite "fair" (Cooper & Richardson, 1986).

One purpose of this article is to examine the relation of job

satisfaction with both dispositional and situational factors in a sample of young adults who, in general, are likely to experience more significant changes in central aspects of their work situations. The greater sample variance in such changes provides a more sensitive test for situational effects. To facilitate comparison with the Staw and Ross results, similar methods are used to some extent, although it should be understood that this similarity does not indicate acceptance of the Staw and Ross methodology (e.g., interpretation of the coefficient on previous job satisfaction as indicative of the importance of stable traits).

Staw and Ross (1985) explicitly mentioned job redesign as an example of a personnel program that may not have the intended effects given their findings. Yet, no measure of job design was included in their study. Thus, a second purpose of the present study is to examine the relation between job complexity (Hackman & Oldham, 1975, 1976, 1980) and job satisfaction. A recent meta-analysis (Loher, Noe, Moeller, & Fitzgerald, 1985) found that only the study by Orpen (1979; $n = 36$) examined the impact of actual changes in job complexity in a field setting. More typically, research in this area has been cross-sectional and has focused on the relation between incumbent self-reports of job complexity and job satisfaction.

However, problems have been raised with the latter research design. First, common methods variance (Schwab & Cummings, 1976; Roberts & Glick, 1981) or priming and consistency effects (Salancik & Pfeffer, 1978) may contribute to the observed relation between self-reports of job complexity and job satisfaction. (See Stone & Gueutal, 1984, however, for evidence contrary to the latter.) Second, job design theory specifies that changes in job complexity will result in changes in job satisfaction. Thus, cross-sectional data have an inherent limitation with respect to testing this theory.

Given these concerns, the present study uses longitudinal data to assess the impact of changes in job complexity on job satisfaction. Moreover, in addition to a self-report measure, an independent complexity measure is derived from the 4th edition of the *Dictionary of Occupational Titles* (DOT; U.S. Department of Labor, 1977). Any relation between the latter measure of complexity and job satisfaction is not likely to be a function of the artifacts discussed by Schwab and Cummings (1976) or by Salancik and Pfeffer (1978). As in the Staw and Ross (1985) research, pay, occupational status, and previous job satisfaction are also examined as possible determinants of current job satisfaction.

Method

Sample

The data are taken from the youth cohort of the NLS, a national probability sample of 12,686 men and women between the ages of 14 and 21 in 1979 (and ages 17–24 in 1982). This sample was interviewed for the first time in 1979, with follow-up interviews conducted annually. The present study uses data from 1979 and 1982. If in either year a person was enrolled in school, was less than 17 years of age, had been working less than 20 hr per week, or had been with the present employer less than 2 months, he or she was excluded from the sample. These restrictions were imposed to ensure the inclusion only of persons with strong labor force attachment. As a consequence, the sample size for the analyses to be reported here, unless noted otherwise, is 809. These

persons were distributed across more than 125 three-digit census occupations.

Measures

As in the Staw and Ross (1985) study, job satisfaction is measured using a single-item global satisfaction measure with four possible levels of response. The anchors range from *dislike it* [the job you have now] *very much* to *like it very much*, slightly different from the global job-satisfaction item used in the Staw and Ross study, in which the anchors range from *highly satisfied* to *highly dissatisfied*. The construct validity of single-item measures is often questioned on at least two grounds: (a) lack of reliability and (b) lack of domain coverage. Nevertheless, in the case of job satisfaction, Scarpello and Campbell (1983) concluded that the single-item global satisfaction item in their study was not unreliable and was the “most inclusive measure of overall job satisfaction” (p. 598).

Also, as in the Staw and Ross (1985) study, employer change, occupational change, pay change, and status (Duncan’s measure of socioeconomic status; Reiss, 1961) change are included as situational variables.

Some evidence suggests that instruments such as the Job Diagnostic Survey and the Job Characteristics Inventory (JCI) measure a unidimensional construct (Aldag, Barr, & Brief, 1981; Drasgow & Miller, 1982). Furthermore, the appropriate label for this construct may be complexity (Gerhart, 1985; Hackman & Oldham, 1980; Stone & Gueutal, 1985).

Thus, the first measure of job complexity is essentially a short form of the JCI (developed by Sims, Szilagyi, & Keller, 1976). Each of the six subscales of the JCI (Variety, Dealing With Others, Autonomy, Feedback, Task Identity, and Friendship Opportunities) is represented by one item that previous research has shown to load strongly on that dimension (see Sims et al., 1976). In addition, an item designed to measure task significance is included. Confirmatory factor analyses supported the use of a single-factor solution. Thus, the seven items were equally weighted and summed to form a scale hereafter referred to as incumbent perceptions of job complexity (IPJC). The internal consistency of this scale as estimated by the Spearman–Brown formula was .72 in 1979, and .79 in 1982. Further psychometric information on this scale can be found in Gerhart (1985).

A second measure of complexity used in the present study was derived by Roos and Treiman (1980) from the 4th edition of the *Dictionary of Occupational Titles* (U.S. Department of Labor, 1977) and is referred to as DOT-complexity in the present study. Sample items include complexity of function in relation to data, required intelligence, and required temperament for repetitive or continuous processes. Psychometric information on the DOT-complexity measure can be found in Cain and Treiman (1981), Cain and Green (1983), and Gerhart (1985).

Cain and Treiman (1981, p. 254) noted that the DOT has the advantage of being based not on worker self-reports but rather “on extensive on-site observation of jobs as they are actually performed and index job content rather than worker characteristics.” It should be understood, however, that this measure represents average levels of complexity for entire occupations rather than jobs.² To the extent that jobs in an occupation differ substantially from one another, error is introduced into the analysis. As a result, the true relation between job complexity and other variables will be underestimated using the DOT-complexity measure.

The DOT-based measure is useful, however, because of the way it complements the IPJC measure. First, in job analysis terms (McCormick, 1976), the DOT-based measure is dominated by items pertaining to personnel requirements, whereas the IPJC measure emphasizes work-oriented activities. Second, the source of information differs (i.e., incumbent vs. nonincumbent). As described earlier, these differences help control the effects of measurement artifacts such as common methods variance.

Table 1
Estimated Reliabilities of Measures

Measure	Reliability of component parts	Correlation between component parts	Reliability of change scores ^a
Job satisfaction	.80		
Previous job satisfaction	.80		
Previous pay	.90		
Current pay	.90		
Previous occupational status	.90		
Current occupational status	.90		
Previous DOT-complexity	.70		
Current DOT-complexity	.70		
Previous IPJC	.72		
Current IPJC	.79		
Δ Pay		.53	.79
Δ Duncan		.58	.76
Δ DOT-complexity		.43	.47
ΔIPJC		.30	.65

Note. DOT = *Dictionary of Occupational Titles*; IPJC = incumbent perceptions of job complexity.

^a Based on the following formula (Guilford, 1954, p. 394):

$$r_{dd} = \frac{r_{jj} + r_{kk} - 2r_{jk}}{2(1 - r_{jk})},$$

where r_{dd} = reliability of the difference $X_j - X_k$, r_{jj} , r_{kk} = respective reliabilities of X_j and X_k , and r_{jk} = intercorrelation of X_j and X_k . For example, in the case of DOT-complexity, $r_{jj} = r_{kk} = .70$, and $r_{jk} = .43$; thus,

$$r_{dd} = \frac{.70 + .70 - 2(.43)}{2(1 - .43)} = .47.$$

Analyses

Because measurement error in either the independent or dependent variables biases estimates of standardized regression coefficients (Kenny, 1979), LISREL (Jöreskog & Sörbom, 1981) was used to correct for the effects of measurement error. Given that change scores can exacerbate measurement error problems, this correction is especially important in the present context. In LISREL, correction for measurement error is made by fixing the error variance, theta delta, to $(1 - r_{xx}) \sigma_x^2$ and by fixing the factor loading, lambda, equal to the square root of the reliability.

The reliability of pay and occupational status measures in surveys like the NLS is approximately .90 (Jencks, 1979). Based on previous research (i.e., Cain & Green, 1983; Kohn & Schooler, 1973; Spenner, 1980), Gerhart (1985) estimated the reliability of the DOT-complexity measure to be approximately .70. This estimate includes both of the following sources of error: (a) interrater differences and (b) the use of an occupation level measure as a proxy for a job level construct. Guilford’s (1954, p. 394) formula was used to estimate the reliabilities of the change scores used. Table 1 summarizes the reliability estimates used to correct for measurement error.

² “Jobs are specific positions within establishments or the economic activities of specific individuals. They entail particular duties and responsibilities and involve the performance of particular tasks in particular settings. . . . Occupations are aggregations of jobs, grouped on the basis of their similarity in content” (Cain & Treiman, 1981, p. 254).

Table 2
Test-Retest Correlations for Global Job Satisfaction as a
Function of Employer Change and Three-Digit
Census Occupation Change

Occupation	Same employer		Changed employer	
	1 ^a	2 ^b	1 ^a	2 ^b
Same				
<i>r</i>	.36	.37	.30	.23
<i>n</i>	139	1,711	90	1,232
Changed				
<i>r</i>	.22	.24	.19	.19
<i>n</i>	234	274	569	1,121

^a Present study (1979 and 1982 data). ^b Corresponding correlations and sample sizes from the Staw and Ross (1985) study.

In addition to correcting for unreliability in the situational variables, job satisfaction was also regressed on the 1979 and 1982 measurements of pay and status, rather than on changes in these variables, because the use of change scores implicitly assumes an arbitrary weighting scheme (i.e., weights of +1 and -1, respectively, for the two component parts of the change score) that may fail to maximize the explanatory power of the regression equation (Cronbach & Furby, 1970; Glansnapp, 1984).

Results

The correlation between 1979 and 1982 satisfaction is .22, somewhat lower than the correlation found by Staw and Ross (1985) between two satisfaction measurements 3 years apart ($r = .32$). Table 2 shows the correlations between satisfaction in 1979 and 1982 as a function of employer and three-digit census occupational change. Table 2 also contains the corresponding results from the Staw and Ross study (using the correlations between 1966 and 1971 satisfaction in their Table 2). The correlations in the two studies are similar despite the fact that the age groups and time intervals differ. In both studies, however, significant differences in consistency exist as a function of the rough proxies for situational change (i.e., change in occupation or employer). These differences are more obvious if one squares the correlations to obtain estimates of variance explained. Both studies, for example, indicate that when neither occupation nor employer changed, the explained variance is 13%–14%. However, when occupation and employer both change, variance explained declines to 4%.

These differences seem to demonstrate that situational changes do, in fact, make a difference even when crudely measured. In addition, however, the similarities across the two studies suggest another important point—the age differences between the samples do not appear to be important *within* these rough classifications. The key, however, is that the proportions within these cells differ strongly between the studies, with a far greater percentage of persons changing occupations and employers in the present sample. This difference explains the smaller overall test-retest correlation found in this study. It also explains the lower test-retest correlations found for pay and status (.53 and .54, respectively, versus .84 for both measures in the Staw and Ross, 1985, study).

Recall that Staw and Ross found little change in R^2 ($\Delta R^2 = .004$) associated with adding changes in pay and status to their equation for job satisfaction. In contrast, the reestimations of the Staw and Ross models using the NLS youth cohort data indicate that changes in pay and status do seem to make a difference in employee job satisfaction (see Table 3). This conclusion is further strengthened by reestimation of the model using (a) the component parts and (b) the corrections for measurement error.

To directly test the effects of changes in job complexity on job satisfaction, models incorporating the two measures of complexity were estimated. These results appear in Table 4. In this case, support for the importance of situational changes is even stronger. Significant increments in R^2 are achieved by adding the situational variables to the equation. When using the separate 1979 and 1982 component scores, the coefficients on the complexity variables approach (DOT-complexity) or exceed (IPJC) the magnitude of the coefficient on previous job satisfaction.

As before, the model was reestimated using the corrected change and component scores. Once again, stronger support for the importance of situational changes was obtained when procedures were used to correct for measurement error. In Table 4, Equation 3b, the coefficient on current DOT-complexity exceeds that for previous job satisfaction. In Equation 5b, the coefficient on current IPJC is more than double that of the coefficient on previous job satisfaction. These results suggest that job complexity may be an important determinant of job satisfaction.

A final issue concerns the possible impact of unreliability in the single-item job-satisfaction measures. Because there is only one item, an internal consistency reliability estimate could not be computed. Nevertheless, it may be that correcting measurement error only in the situational variables stacks the deck in favor of finding significant situational effects. This issue can be

Table 3
Job Satisfaction as a Function of Previous Job Satisfaction,
Pay, and Occupational Status

Variable	Uncorrected ^a			Corrected ^b	
	1a	2a	3a	2b	3b
Previous job satisfaction	.26**	.26**	.26**	.27**	.26**
Previous pay			-.09*		-.11**
Current pay			.07*		.09*
Previous occupational status			-.05		-.07
Current occupational status			.15**		.17**
Δ Pay		.06*		.07*	
Δ Status		.09**		.11**	
R^2	.066	.078	.087	.082	.092

Note. Different equation numbers (1, 2, 3) indicate inclusion of different sets of predictors. Different equation letters (a, b) indicate different assumptions about measure reliabilities. Entries in columns are standardized partial regression coefficients.

^a Not corrected for measurement error. ^b Corrected for measurement error in situational variables only.

* $p < .05$, one-tailed. ** $p < .01$, one-tailed.

Table 4
Job Satisfaction as a Function of Previous Job Satisfaction, Pay, IPJC, and DOT-Complexity

Variable	Uncorrected ^a					Corrected ^b			
	1a	2a	3a	4a	5a	2b	3b	4b	5b
Previous job satisfaction	.26**	.27**	.26**	.31**	.20***	.29**	.27**	.34**	.21**
Previous pay			-.09*		-.06*		-.10*		-.06
Current pay			.06		.01		.07		-.00
Previous DOT-complexity			-.05				-.14*		
Current DOT-complexity			.18**				.28**		
Previous IPJC					-.04				-.09*
Current IPJC				-.04	.39**				.47**
Δ Pay		.06*				.07*		.03	
Δ DOT-complexity		.13**				.20**			
Δ IPJC				.28**				.36**	
R ²	.066	.087	.100	.145	.212	.108	.122	.189	.257

Note. Different equation numbers (1, 2, 3, 4, 5) indicate inclusion of different sets of predictors. Different equation letters (a, b) indicate different assumptions about measure reliabilities. Entries in columns are standardized partial regression coefficients. IPJC = incumbent perceptions of job complexity; DOT = *Dictionary of Occupational Titles*.

^a Not corrected for measurement error. ^b Corrected for measurement error in situational variables only.

* $p < .05$, one-tailed. ** $p < .01$, one-tailed.

addressed in two ways. First, note that both Tables 3 and 4 indicate that even without correction for measurement error, the situational factors have a significant effect on job satisfaction. Nevertheless, a second strategy was to reestimate the corrected models using arbitrary reliability estimates of .80 and .60 for both satisfaction measures to determine if the findings would be altered. As Tables 5 and 6 demonstrate, the relative importance of the situational factors is largely unaffected by this unreliability correction.

Discussion

One purpose of this study was to assess recent suggestions that traits may be more important determinants of job satisfaction than previously believed and that personnel programs designed to change the work environment may be rendered ineffective in many cases because of the impact of these stable

traits. Results of the present study, however, indicate that pay, status, and job complexity added explanatory power to an equation predicting job satisfaction, controlling for earlier job satisfaction. The job complexity measures, in particular, were strong predictors.

These findings differ strongly from the Staw and Ross (1985) study that found little predictive power for situational factors. Perhaps the most straightforward explanation for these conflicting findings is that the present study used a sample that experienced more significant variance in changes in job attributes over time. This sample characteristic seems desirable because personnel programs such as job design entail changes in job attributes. In addition, the present research directly examined changes in job complexity, the core construct of Hackman and Oldham's (1975, 1976) job design model. The focus on estimating relations corrected for measurement error also contributed to stronger estimated situational effects.

Table 5
Robustness of Results to Varying Assumed Levels of Job Satisfaction Reliability: Occupational Status

Variable	Reliability ^a = .80			Reliability ^a = .60		
	1c	2c	3c	1d	2d	3d
Previous job satisfaction	.32**	.33**	.33**	.43**	.45**	.45**
Previous pay			-.12**			-.14*
Current pay			.09*			.09
Previous occupational status			-.10*			-.15*
Current occupational status			.20**			.25**
Δ Pay		.08*			.08	
Δ Status		.13**			.17**	
R ²	.102	.125	.137	.182	.216	.230

Note. Situational variables are corrected for measurement error. Column entries are standardized partial regression coefficients. Different equation numbers (1, 2, 3) indicate inclusion of different sets of predictors. Different equation letters (c, d) indicate different assumptions about measure reliabilities.

^a Assumed reliability of previous and current job satisfaction.

* $p < .05$, one-tailed. ** $p < .01$, one-tailed.

Table 6

Robustness of Results to Varying Assumed Levels of Job Satisfaction Reliability: Job Complexity

Variable	Reliability ^a = .80					Reliability ^a = .60				
	1c	2c	3c	4c	5c	1d	2d	3d	4d	5d
Previous job satisfaction	.32**	.37**	.35**	.44**	.28**	.43**	.50**	.48**	.60**	.45**
Previous pay			-.11**		-.06			-.12**		-.06
Current pay			.07		-.01			.06		-.01
Previous DOT-complexity			-.18**					-.26**		
Current DOT-complexity			.32**					.40**		
Previous IPJC					-.15**					-.28**
Current IPJC					.52**					.60**
Δ Pay		.07		.03			.06		.01	
Δ DOT-complexity		.24**					.31**			
Δ IPJC				.43**					.55**	
R ²	.102	.162	.176	.274	.336	.182	.276	.289	.452	.490

Note. Situational variables are corrected for measurement error. Column entries are standardized partial regression coefficients. Different equation numbers (1, 2, 3, 4, 5) indicate inclusion of different sets of predictors. Different equation letters (c, d) indicate different assumptions about measure reliabilities. DOT = *Dictionary of Occupational Titles*; IPJC = incumbent perceptions of job complexity.

^a Assumed reliability of previous and current job satisfaction.

* $p < .05$, one-tailed. ** $p < .01$, one-tailed.

These findings are potentially important for the job design area because of several methodological strengths. First, longitudinal data were used. Thus, the effect of changes in levels of complexity on job satisfaction could be examined. Second, the research was conducted in a field setting and included a wide range of occupations. Therefore, the lack of external validity sometimes attributed to laboratory research using students may have been less of a problem (see Stone, 1986, however, for evidence supporting the external validity of laboratory research in the job design area). Third, both measures of complexity were related to job satisfaction. Given that the two measurement methods for complexity were quite different, it is unlikely that the observed complexity-satisfaction relation can be explained in terms of measurement artifacts such as priming and consistency effects or common methods variance.

Despite these advantages, note that the changes experienced by workers in this sample may be qualitatively different from the changes experienced as the result of an intervention in a single organization. Moreover, there may be unmeasured changes in other situational factors confounded with changes in complexity that also influence satisfaction. If true, the coefficients on the complexity variables would be biased. To examine this possibility, several additional situational factors³ were added to Table 4, Equations 3a and 5a. The addition of these variables, however, changed the coefficients on current IPJC and current DOT-complexity by less than 10%, suggesting that the estimated effects of complexity may, in fact, be quite robust.

Similar to Staw and Ross's (1985) findings, I found that previous job satisfaction predicted current job satisfaction. One interpretation of this predictive power is to follow the reasoning of Staw and Ross and attribute it to the importance of traits that remain stable over time and across situations.

There are two problems with this explanation, however. First, to the extent that important situational variables are omitted or poorly measured, the relative predictive power of previous job satisfaction will appear greater. Second, and closely related, no specific individual variables were specified by Staw and Ross

as accounting for the (moderate) stability in job satisfaction. Instead, any unexplained stability in job satisfaction was attributed to unspecified traits. The theoretical rationale for such an attribution is not at all clear, as was later acknowledged by Staw et al. (1986): "Such consistency data do not, however, constitute a dispositional theory of attitudes, since they have little to say about *why* individuals may show stability in job satisfaction" (p. 60). As a consequence, Staw et al. attempted to better define a specific trait (i.e., affective disposition).

It may be more accurate to interpret the predictive power of previous job satisfaction in the Staw and Ross (1985) study as indicative of some degree of stability in *both* traits and job attributes given that previous job satisfaction is itself a function of both types of factors (Locke, 1969). Thus, an observed relation between previous and current job satisfaction should perhaps be viewed as an upper bound on the total effect of traits on job satisfaction.

Finally, from an applied view, an impact of traits on global job satisfaction may have little relevance for at least two reasons. First, organizations often focus their efforts on improving specific aspects of the work situation with which employees are dissatisfied. Specific facet satisfaction measures may better reflect changes in relevant situational factors because of the more precise referent. Although the NLS has sparse information on facet satisfaction, it did ask respondents how true the following statement was of their job: The pay is good. The R^2 obtained by predicting this rough measure of pay satisfaction in 1982 with the corresponding 1979 measure was .07. Adding pay change

³ The 1979 and 1982 measures for the following variables were added to the equations: (a) working conditions, (b) physical demands, (c) required motor skills, (d) firm tenure, (e) weekly hours worked, (f) presence or absence of health insurance, (g) presence or absence of a paid vacation benefit, (h) presence or absence of life insurance, and (i) Duncan's measure of occupational status (the same as that included in Table 3). Inclusion of Duncan's measure of occupational status helps control for any influence of occupational progression among young adults.

to the equation, however, increased the R^2 to .20 (without any correction for measurement error). This finding provides tentative support for the proposition that specific facet (vs. global) satisfaction measures may be more responsive to changes in situational factors.

Second, one may simply not care if satisfaction is correlated over time due to stable traits. More relevant in many cases will be the question of whether the average level of job satisfaction of workers can be increased through the use of some program. If, as the present results suggest, the level does increase in response to certain situational changes, the question of whether workers maintain their relative positions over time (as assessed by correlational methods) may not be important.⁴

In conclusion, the present research does not support Staw and Ross's (1985) conclusion that attitudinal consistency will equal or exceed the effects of situational factors in most contexts. Support for the importance of situational factors such as job complexity was found. In contrast, direct evidence for the importance of traits as determinants of job satisfaction was not obtained in either study. An adequate test for the impact of traits requires the specification and testing of models containing specific trait measures.

Implications

I suggest the following important practical implications. First, there is little evidence that "Many situational interventions may be prone to failure because they must contend with attitudinal consistency" (Staw & Ross, 1985, p. 478; see also Staw et al., 1986). The present results demonstrate that changes in situational factors such as job complexity and pay may have an important impact even on global job satisfaction. This impact may be greater for specific facets of satisfaction that personnel programs often target.

Second, even if there is stability in the relative satisfaction of workers over time, the overall level of satisfaction may still be increased by well-designed personnel programs. Therefore, the practical significance of attitudinal consistency may be minimal unless such consistency is shown to place an important constraint on changes in job-satisfaction levels.

Finally, until more compelling evidence for the impact of stable traits on job satisfaction is found, personnel selection based on traits (e.g., propensity to be satisfied) may be premature. It is important to remember that even conceptually stable traits such as intellectual flexibility have been found to change in response to situational factors like job complexity (Kohn & Schooler, 1973, 1982). To the extent that traits lack stability, their predictive validity is diminished. The longstanding conclusion that personality traits have suspect predictive validity (Guion & Gottier, 1965) may still apply.

⁴ Statistical methods requiring longitudinal data (e.g., first-differencing models) are available that yield unbiased estimates of changing situational factors even in the presence of unmeasured, but stable, trait factors (e.g., Gerhart, 1985; Mundlak, 1978).

References

- Aldag, R. J., Barr, S. H., & Brief, A. P. (1981). Measurement of perceived task characteristics. *Psychological Bulletin*, 90, 415-431.
- Cain, P. S., & Green, B. F. (1983). Reliabilities of selected ratings available from the Dictionary of Occupational Titles. *Journal of Applied Psychology*, 68, 155-165.
- Cain, P. S., & Treiman, D. J. (1981). The Dictionary of Occupational Titles as a source of occupational data. *American Sociological Review*, 46, 253-278.
- Cronbach, L. J., & Furby, L. (1970). How we should measure "change"—or should we? *Psychological Bulletin*, 74, 68-80.
- Cooper, W. H., & Richardson, A. J. (1986). Unfair comparisons. *Journal of Applied Psychology*, 71, 179-184.
- Drasgow, F. P., & Miller, H. E. (1982). Psychometric and substantive issues in scale construction and validation. *Journal of Applied Psychology*, 67, 268-279.
- Duncan, G. J., & Hill, D. H. (1985). An investigation of the extent and consequences of measurement and error in labor-economic survey data. *Journal of Labor Economics*, 3, 508-532.
- Gerhart, B. (1985). *Sources of variance in perceptions of job complexity*. Unpublished doctoral dissertation, University of Wisconsin—Madison.
- Glansnapp, D. R. (1984). Change scores and regression suppressor conditions. *Educational and Psychological Measurement*, 44, 851-867.
- Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.
- Guion, R. M., & Gottier, R. F. (1965). Validity of personality measures in personnel selection. *Personnel Psychology*, 31, 841-852.
- Hackman, J. R., & Oldham, G. R. (1975). Development of the Job Diagnostic Survey. *Journal of Applied Psychology*, 60, 159-170.
- Hackman, J. R., & Oldham, G. R. (1976). Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance*, 16, 250-279.
- Hackman, J. R., & Oldham, G. R. (1980). *Work redesign*. Reading, MA: Addison-Wesley.
- Jencks, C. J. (1979). *Who gets ahead?* New York: Basic Books.
- Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V*. Uppsala, Sweden: University of Uppsala.
- Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley.
- Kohn, M. L., & Schooler, C. (1973). Occupational experience and psychological functioning: An assessment of reciprocal effects. *American Sociological Review*, 38, 197-118.
- Kohn, M. L., & Schooler, C. (1982). Job conditions and personality: A longitudinal assessment of their reciprocal effects. *American Journal of Sociology*, 87, 1257-1286.
- Locke, E. A. (1969). What is job satisfaction? *Organizational Behavior and Human Performance*, 4, 309-336.
- Loher, B. T., Noe, R. A., Moeller, N. L., & Fitzgerald, M. P. (1985). A meta-analysis of the relation of job characteristics to job satisfaction. *Journal of Applied Psychology*, 70, 280-289.
- McCormick, E. J. (1976). Job and task analysis. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 651-696). Chicago: Rand-McNally.
- Mitchell, T. R. (1979). Organizational behavior. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual review of psychology* (Vol. 30, pp. 243-281). Palo Alto, CA: Annual Reviews.
- Mundlak, Y. (1978). On the pooling of time series and cross-sectional data. *Econometrica*, 46, 69-85.
- Orpen, C. (1979). The effects of job enrichment on employee satisfaction, motivation, involvement, and performance: A field experiment. *Human Relations*, 32, 189-217.
- Pulakos, E. D., & Schmitt, N. (1983). A longitudinal study of a valence model approach for the prediction of job satisfaction. *Journal of Applied Psychology*, 68, 307-312.
- Reiss, A. (1961). *Occupations and social status*. New York: Free Press.
- Roberts, J., & Glick, W. (1981). The job characteristics approach to task design: A critical review. *Journal of Applied Psychology*, 66, 193-217.

- Roos, P. A., & Treiman, D. J. (1980). Worker functions and worker traits for the 1970 U.S. census classification. In A. R. Miller, D. J. Treiman, P. S. Cain, & P. A. Roos (Eds.), *Work, jobs, and occupations: A critical review of the Dictionary of Occupational Titles* (Appendix F). Washington, DC: National Academy Press.
- Salancik, G. R., & Pfeffer, J. (1978). A social information processing approach to job attitudes and task design. *Administrative Science Quarterly*, 23, 224-253.
- Scarpello, V., & Campbell, J. P. (1983). Job satisfaction: Are all the parts there? *Personnel Psychology*, 36, 577-600.
- Schwab, D. P., & Cummings, L. L. (1976). Impact of task scope on employee productivity: An evaluation using expectancy theory. *Academy of Management Review*, 1, 23-35.
- Sims, H., Szilagyi, A., & Keller, R. (1976). The measurement of job characteristics. *Academy of Management Journal*, 19, 195-212.
- Spencer, K. I. (1980). Occupational characteristics and classification systems: New use of the Dictionary of Occupational Titles in social research. *Sociological Methods and Research*, 9, 239-264.
- Staw, B. M., Bell, N. E., & Clausen, J. A. (1986). The dispositional approach to job attitudes: A lifetime longitudinal test. *Administrative Science Quarterly*, 31, 56-77.
- Staw, B. M., & Ross, J. (1985). Stability in the midst of change: A dispositional approach to job attitudes. *Journal of Applied Psychology*, 70, 469-480.
- Stone, E. F. (1986). Job scope-job satisfaction and job scope-job performance relationships. In E. A. Locke (Ed.), *Generalizing from laboratory to field settings: Research findings from industrial-organizational psychology, organizational behavior and human resource management* (pp. 189-206). Lexington, MA: Lexington Books.
- Stone, E. F., & Gueutal, H. (1984). On the premature death of need satisfaction models: An investigation of Salancik and Pfeffer's views on priming and consistency artifacts. *Journal of Management*, 10, 237-258.
- Stone, E. F., & Gueutal, H. (1985). An empirical derivation of the dimensions along which characteristics of jobs are perceived. *Academy of Management Journal*, 28, 376-396.
- U.S. Department of Labor. (1977). *Dictionary of Occupational Titles* (4th ed.). Washington, DC: U.S. Government Printing Office.
- Weiss, H. M., & Adler, J. (1984). Personality and organizational behavior. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. 6, pp. 1-50). Greenwich, CT: JAI Press.

Received March 28, 1986

Revision received February 2, 1987

Accepted January 8, 1987 ■

Correction to Earley et al.

In the article "Task Planning and Energy Expended: Exploration of How Goals Influence Performance," by P. Christopher Earley, Pauline Wojnarowski, and William Prest (*Journal of Applied Psychology*, 1987, Vol. 72, No. 1, 107-114), Table 1 on page 109 appeared with reversed labels for the high and low conditions. The corrected Table 1 is as follows:

Table 1
Means, Standard Deviations, and Intercorrelations of
Performance, Planning, and Energy Expended for
Goal and Information in Study 1

Measure and information	Goal			
	"Do best"		Specific	
	M	SD	M	SD
Performance				
Low	8.26	1.98	20.25	6.72
High	15.32	2.26	25.14	6.56
Planning				
Low	1.78	0.49	3.23	0.52
High	2.57	0.73	3.39	0.85
Energy expended				
Low	2.61	0.60	3.24	0.63
High	3.35	0.53	3.76	0.59

^a n = 18/cell. ^b Dummy-coded 0, 1.

Unemployment, Job Satisfaction, and Employee Turnover: A Meta-Analytic Test of the Muchinsky Model

Jeanne M. Carsten and Paul E. Spector
University of South Florida

We conducted a meta-analysis to determine the relation between satisfaction–turnover correlations across studies and unemployment rates at the time those studies were conducted. On the basis of theoretical work by Muchinsky and Morrow (1980), we hypothesized that low relations would be found in studies conducted during times of high unemployment and limited employment opportunity, and high relations would be found in studies conducted during times of low unemployment and expanded opportunity. Results supported the hypothesis; correlations were found that ranged from $-.18$ to $-.52$ between unemployment rates and the magnitude of satisfaction–turnover relations across studies. A similar analysis was conducted for the relation between intention to quit and turnover. The correlations between the intention–turnover relation and unemployment were similar in magnitude to the corresponding satisfaction correlations, indicating that the behavioral–intention–turnover relation is also moderated by economic alternatives. Severe methodological problems with a similar study, which indicated the opposite results (Shikiar & Freudenber, 1982), are discussed.

The relation between job satisfaction and voluntary employee turnover has been heavily researched (see reviews by Mobley, Griffeth, Hand, & Meglino, 1979; Muchinsky & Tuttle, 1979; Porter & Steers, 1973; Price, 1977). The general conclusion of these reviews is that there is a moderate correlation between job satisfaction and turnover; that is, dissatisfied employees are more likely to quit their jobs than are their satisfied colleagues. There is considerable variation in correlations across studies but they are usually less than $.40$ (Locke, 1976), thus leaving the majority of variance associated with turnover unexplained.

A larger proportion of turnover variance has been explained by behavioral intentions. Expressed intention to leave was suggested as a cognitive intermediary between job dissatisfaction and employee turnover by Mobley (1977) and Porter and Steers (1973). Subsequent research has supported this model, indicating that behavioral intentions are a stronger predictor of employee turnover than is job satisfaction (Atchison & Lefferts, 1972; Kraut, 1975; Michaels & Spector, 1982; Mobley, Horner, & Hollingsworth, 1978; Parasuraman, 1982; Waters, Roach, & Waters, 1976), but again the magnitude of the correlations is quite variable.

Muchinsky and Morrow (1980) proposed a turnover model that may explain the variation among correlations. They hypothesized three major sets of turnover determinants: economic opportunity factors, individual factors, and work-related fac-

tors. Economic opportunity factors, including both local and national unemployment, were said to have the strongest impact on turnover. Job dissatisfaction was considered a precursor of turnover, but its effect is moderated by economic factors. During periods of high unemployment and low opportunity for alternative employment, relatively few individuals will quit and the correlation between satisfaction and turnover will be low. During low unemployment and high opportunity, more of the dissatisfied employees will quit and the satisfaction–turnover correlation will be higher. Thus, the economy acts as a releaser, allowing satisfaction to best predict turnover during periods of high economic opportunity (Hulin, Roznowski, & Hachiya, 1985).

The support for the predominance of economic factors in the model has been reviewed by Muchinsky and Tuttle (1979) and is again cited by Muchinsky and Morrow (1980). Labor turnover has been shown to be strongly associated with unemployment (Behrend, 1951; Crowther, 1957; Eagly, 1965). For example, Eagly obtained a correlation coefficient of $-.84$ between quit rates and the national unemployment rate in the United States from 1931 to 1962.

A recent meta-analytic study by Shikiar and Freudenber (1982) examined unemployment rates as a moderator of the job satisfaction–turnover relation. These authors obtained a $.39$ correlation coefficient in an archival study correlating unemployment rates with the satisfaction–turnover correlations from past studies. Their results indicated that the relation between job satisfaction and turnover was stronger during periods of relatively high unemployment, or fewer alternatives, than during periods of relatively low unemployment, or greater alternatives (Shikiar & Freudenber, 1982). This finding directly contradicts the Muchinsky and Morrow model.

There were severe methodological problems, however, which make uninterpretable the Shikiar and Freudenber findings.

This article is based on the first author's master's thesis under the direction of the second author, submitted in partial fulfillment of requirements for the Master of Arts degree at the University of South Florida.

Correspondence concerning this article should be addressed to Paul Spector, Department of Psychology, University of South Florida, Tampa, Florida 33620.

The most significant problem was their method of determining corresponding unemployment data for each study in the analysis. Obviously, it was essential that the unemployment rate was taken from the same time period as the turnover data. Shikiar and Freudenberg assumed that the data for each study were collected 2 years before the publication date of the article. As we will show, this was a faulty assumption.

A second problem involved their study selection procedures. First, most of the studies with the higher job satisfaction–turnover correlations were based on aggregate data; for example, Ley (1966) obtained a job satisfaction–turnover correlation of .79 when sampling work teams rather than individual employees. The use of aggregate data may inflate correlation coefficients over individual-level analysis (Langbein & Lichtman, 1978), so the two levels of data should not be mixed. Second, when an overall job satisfaction–turnover coefficient was not available, they averaged the coefficients of the satisfaction subscales and used that as an overall index. This procedure may not provide an accurate index of the overall satisfaction–turnover correlation because all possible satisfaction dimensions were not sampled. The use of a single dimension is not appropriate when the question of interest concerns overall job satisfaction. Third, Shikiar and Freudenberg (1982) included cases that used variables other than job satisfaction. A few of the studies had investigated organizational commitment and its relation to turnover. Such cases should not have been included because these variables are conceptually different and are not related to employee turnover to the same degree. Of particular concern is the fact that organizational commitment scales tend to be confounded with behavioral intentions, including items concerning how long the individual intends to remain with the organization. Finally, the authors omitted nonsignificant correlations from their sample, thus restricting the range of correlations studied.

The purpose of this study was to replicate the Shikiar and Freudenberg (1982) study and correct the methodological problems that occurred in the original work. A meta-analysis was conducted on available satisfaction–turnover studies, including the appropriate studies from Shikiar and Freudenberg's sample and additional studies that were located. The authors of each article were contacted to ascertain the exact time period of data collection. Those cases in which the information could not be acquired were eliminated from the sample. In addition, the occupation and industry of each sample and the state(s) in which data were collected were ascertained, to allow for the use of more precise and locally relevant unemployment rates. Finally, the analysis was extended to include the moderating effect of unemployment rates on the behavioral-intention–turnover relation, as well as the satisfaction–turnover relation.

This study was a test of the Muchinsky and Morrow (1980) model. Despite Shikiar and Freudenberg's contrary findings, we hypothesized that as the number of employment opportunities decreased, the job satisfaction–turnover relation would become weaker. Intention to quit was not considered by either Shikiar and Freudenberg (1982) or Muchinsky and Morrow. A hypothesis for the behavioral-intention–turnover relation was based on Mobley's (1977) microanalytic treatment of the behavioral-intention–turnover linkages. Mobley (1977) suggested that these intentions occur after an individual has assessed the job market,

and thus the intention–turnover relation should remain constant across changes in unemployment. We hypothesized, therefore, that the behavioral-intention–turnover relation would not be affected by employment opportunities.

Method

Studies

Studies were identified for use in this research by three methods. First, the bibliographies of comprehensive reviews were used, specifically, Mobley et al. (1979), Muchinsky and Tuttle (1979), Porter and Steers (1973), Shikiar and Freudenberg (1982), and Steel and Ovalle (1984). More recent literature was acquired through the *Psychological Abstracts* (American Psychological Association, 1977–1984) and by scanning the table of contents of the most current, salient journals. The criteria for use in this analysis were (a) overall job satisfaction must have been assessed; (b) job satisfaction–turnover or behavioral-intention–turnover relation, or both, must either have been expressed as a correlation coefficient or have been convertible to a correlation coefficient; and (c) the exact year and, if possible, the quarter of data collection must have been ascertained. In addition, to be included in the finer level analysis the particular locale (e.g., state or region of the study) must have been ascertained, or the particular occupation or industry of the sample, or both, must have been ascertained.

The dates of data collection in the studies ranged from 1947 to 1983. Data were collected throughout the United States, Canada, and parts of Europe, and across many occupations and industries (business, health care, service, clerical, and defense). The reference list of included studies is presented in Appendix A.

Measures

The unemployment rates were obtained from the *Statistical Abstract of the United States* (Annual) and *Historical Statistics of the United States* (Dodd & Dodd, 1973). The tables in these volumes included the necessary international rates as well. The unemployment rates used included national, industrial, occupational, and state unemployment rates, and were averaged across the months of the year or years during which satisfaction, behavioral intention, and turnover data were collected. If, in a particular study, data were collected across occupations or states, the individual unemployment rates were weighted and averaged. If more than three occupations, industries, or states were sampled, that particular case was not included in the respective analyses. For instance, military studies that sampled across many states, occupations, and industries were only included in the national unemployment rate analysis.

Procedure

An initial pool of potential studies was identified from the sources previously outlined. The published studies were examined to acquire as much of the necessary information as possible and to screen out those that failed to meet the criteria. The senior authors of the remaining studies were contacted by letter to ascertain where and when the study took place, and which occupation and industry were sampled. The location was expressed as the state or region in which data were collected. The time span was expressed as the month(s) of the year(s) during which the entire process of data collection occurred. If the senior author could not be reached, an attempt was made to contact the other authors of the particular article. A follow-up telephone call was made if there was no response to the letter. If none of the authors could be contacted, or if the requested information was somehow unobtainable, the study was omitted from the sample.

Analysis

The analysis was based on the procedures outlined by Hunter, Schmidt, and Jackson (1982), and Rosenthal (1984). Satisfaction–turnover and intention–turnover relations expressed as student *ts* or chi-squares were first transformed to *rs*. To normalize the distributions, Fisher’s *r* to *z* conversions were used to transform both the job satisfaction–turnover correlation coefficients and the intention-to-quit–turnover correlation coefficients for each study. The *z* statistics were then correlated with the national, industrial, occupational, and state unemployment rates. For this analysis, the signs of the satisfaction–turnover correlations were reversed.

The analysis was repeated adjusting for skewness in the dichotomous turnover distribution. Increases in skewness of turnover artifactually lower the ceiling on the satisfaction and intention correlations with turnover. This statistical artifact was eliminated by partialing out a measure of the skewness effect from satisfaction–turnover and intention–turnover correlations with the unemployment indices. The measure of this effect of skewness was computed using the following formula:

$$\text{Log}(.50 - |.50 - \text{Study Quit Rate}|).$$

Results

A list of the studies with sample sizes, time period, unemployment rates, satisfaction–turnover correlations, and behavioral-intention–turnover correlations is presented in Appendix B. There were 47 cases (satisfaction and intentions, see Appendix A), and a total sample size of 19,828 individuals. The average sample size per case was 413 (*SD* = 863), and the sample sizes ranged from 42 to 5,780. The mean study quit rate was 31.9% (*SD* = 19%), with values ranging from 6% to 85%. This quit rate was inflated by the military studies, which had a mean of 49.8% (*SD* = 26%), with values ranging from 8% to 85%. The mean quit rate of the civilian studies was 24.9% (*SD* = 10%).

Table 1 summarizes the unemployment rates associated with the studies. Included in the table for national, industrial, occupational, and state rates are means, standard deviations, and ranges. As can be seen, rates ranged from 1.6%, generally associated with economic prosperity, to 10.6%, generally associated with recessions and relatively poor economic times in western society.

As can be seen in Table 2, the mean unweighted correlation across studies between job satisfaction and turnover was $-.23$ (dissatisfied most likely to quit), whereas the mean unweighted correlation between behavioral intentions and turnover was $+.38$ (greater intent associated with quitting). Using procedures described in Hunter, Schmidt, and Jackson (1982), the mean correlations were adjusted for attenuation due to unreliability.

Table 1
Unemployment Rates

Unemployment rate	<i>M</i>	<i>SD</i>	Low	High	No. of cases
National	6.01	1.61	3.15	9.70	47
Industrial	5.14	1.87	1.60	8.00	28
Occupational	6.06	2.34	2.80	10.60	27
State	5.95	1.98	2.90	10.00	28

Table 2
Descriptive Statistics

Variable	Job satisfaction and turnover	Behavioral intention and turnover
No. of cases	39	29
Total sample size	12,045	13,711
<i>M</i> sample size	309	473
Range of sample sizes	42–1852	42–5780
<i>M</i> uncorrected <i>r</i>	$-.23$	$+.38$
<i>M r</i> adjusted for attenuation due to unreliability	$-.26$	$+.47$
<i>M r</i> weighted by sample size	$-.22$	$+.26$
<i>M r</i> weighted and adjusted for unreliability	$-.24$	$+.32$
Confidence interval	$-.46$ – $-.02$	$-.08$ – $+.61$
Variance of <i>M</i> weighted <i>r</i>	.015	.032
Variance corrected for error	.012	.030
χ^2 for homogeneity of <i>r</i>	204, $p < .001$	625, $p < .001$

The estimate for reliability was averaged across those studies in which it was reported, to obtain a mean reliability. Satisfaction had a reliability of .85 and intention of .66. Because turnover was an objective measure, its reliability was assumed to be 1.0. After this adjustment, the mean satisfaction–turnover correlation was $-.26$ and the mean behavioral-intention–turnover correlation was $+.47$.

The mean correlation coefficients were also computed, with each study weighted by sample size. The satisfaction–turnover coefficient was essentially unchanged ($r = -.22$), but the intention–turnover coefficient was substantially decreased ($r = +.26$). Further examination revealed that two of the weakest intention–turnover correlations had extremely large sample sizes (5,780 and 1,445). One of these cases is associated with a relatively low unemployment rate (4.25), whereas the other is associated with a relatively high unemployment rate (7.7).

The mean weighted correlation coefficients for satisfaction–turnover and intention–turnover were adjusted for unreliability, yielding coefficients of $-.24$ and $+.32$, respectively. Confidence intervals of 95% around the weighted mean for satisfaction–turnover and intention–turnover were calculated. For satisfaction–turnover and intention–turnover, the variance of the observed correlations was calculated, along with the variance due to sampling error. The observed variance adjusted for sampling error was reduced very little. Sampling error accounted for only 21% of the variance in satisfaction correlations and 3% of the variance in intention correlations. This procedure, according to Hunter et al. (1982), indicates that the differences in correlation coefficients across studies is largely due to one or more moderator variables.¹ Chi-square tests for the homogeneity of a set of correlation coefficients were found to be significant for both the satisfaction–turnover, $\chi^2(38, N = 12,045) = 204, p < .001$, and the intention–turnover, $\chi^2(28, N = 13,711) = 625,$

¹ The accuracy of this adjustment formula has recently been questioned by Spector and Levine (1987). The results of a Monte Carlo study conducted by these authors suggested that the variance formula may overadjust for the artifact of error.

Table 3
Correlations Between Job-Satisfaction–Turnover and Behavioral-Intention–Turnover Indices and Unemployment Rates

Measure and sample	Job satisfaction and turnover			Behavioral intention and turnover		
	<i>r</i>	Partial <i>r</i> ^a	<i>n</i>	<i>r</i>	Partial <i>r</i> ^a	<i>n</i>
Unemployment rate						
National						
Total	–.34	–.32	39	–.30	–.28	29
Civilian	–.48	–.50	30	–.29	–.32	18
Military	–.20	–.40	9	–.03	–.11	11
Industrial						
Total (all civilian)	–.38	–.40	24	–.36	–.40	16
Occupational						
Total (all civilian)	–.52	–.57	23	–.28	–.33	15
State						
Total	–.18	–.10	26	–.36	–.30	16
Civilian	–.35	–.39	22	–.46	–.53	13
Time period of data collection						
Total	–.51	—	39	–.07	—	29
Civilian	–.24	—	30	–.17	—	18
Military	–.84	—	9	–.35	—	11

^a Skewness of the turnover distribution has been partialled out.

$p < .001$, coefficients, again suggesting that sampling error does not fully account for variations in correlations.

The hypotheses of this study were tested with analyses summarized in Table 3. The total sample unadjusted coefficients were computed first. As can be seen, the correlations between job satisfaction–turnover and unemployment ranged from $-.18$ to $-.52$, indicating that as the unemployment rate increased, the job satisfaction–turnover relation decreased. The correlations between the behavioral-intention–turnover index and the unemployment rates were all in the same direction and of similar magnitude (ranging from $-.28$ to $-.36$). The civilian and military data were analyzed separately for both satisfaction–turnover and intention–turnover. For the civilian data, all correlations increased, except the intention–turnover and national unemployment coefficient, which remained about the same. The separate analysis of the military data resulted in decreased correlation coefficients (see Table 3).

The analyses were repeated using the skewness adjustment previously described. This adjustment had only minor effects on the magnitude of most correlation coefficients, as shown in Table 3. The largest change was for satisfaction–turnover and national rate in the military sample ($r = -.20$ vs. $-.40$, for zero order and partial, respectively). For the civilian sample, the occupational rate was most strongly correlated with satisfaction–turnover, followed by national, industrial, and state rates. The intention–turnover index was most strongly correlated with the state rate, followed by industrial, national, and occupational rates. For the military sample, the correlation between satisfaction–turnover and national unemployment was $-.40$, whereas the correlation between intention–turnover and national unemployment rate was $-.11$.

The mean time period between data collection and the publication of the reviewed cases was 3.17 years ($SD = 1.73$ years), with a range from 1 to 8 years. The mean time period during which turnover data were collected was 17 months ($SD = 20.0$

months), with a range from 3 months to 8½ years. The correlation between satisfaction–turnover and the data collection time period was $-.51$ for the total sample, indicating that the job-satisfaction–turnover relation is stronger when the time span of turnover assessment, which begins with the job satisfaction measure, is relatively short. Subsequent analyses indicated that this high correlation coefficient between time period and satisfaction–turnover may be attributed to the military studies ($r = -.84$, $M = 28.0$ months, $SD = 35.0$ months) rather than the civilian studies ($r = -.24$, $M = 13.1$ months, $SD = 7.2$ months). The correlation between intention–turnover and data collection time period was $-.07$ for the total sample, $-.17$ for the civilian sample, and $-.35$ for the military sample.

Discussion

The results of the analysis support the Muchinsky and Morrow (1980) model. This model predicts that the relation between job satisfaction and turnover will be strong during periods of low unemployment (economic prosperity) and weak during periods of high unemployment (economic hardship). These results indicate that the job satisfaction–turnover relation became weaker as the employment level (available alternatives) decreased.

This finding is opposite to the results of Shikiar and Freudenberg (1982). These authors obtained a moderate, positive relation between the job satisfaction–turnover index and unemployment rate. They suggested that the job satisfaction–turnover relation is stronger during periods of fewer employment alternatives because the individuals who do quit at this time are extremely dissatisfied. Apparently, Shikiar and Freudenberg's methodological problems produced distorted results. Their positive correlation is due largely to the two strongest satisfaction–turnover coefficients. These coefficients were both based on aggregate data and should not have been analyzed with the

individual data. When the data were reanalyzed without these two cases, a correlation of zero was obtained. In addition, we found that most studies took longer than their assumed 2 years to get into print, invalidating most of Shikiar and Freudenberg's unemployment statistics.

Our data support the more popular interpretation of the effect of employment alternatives on the job satisfaction–turnover relation. Specifically, that during periods of relatively low unemployment the satisfaction–turnover relation will be strong, and during periods of relatively high unemployment this relation will be weakened. A simple rationale behind this hypothesis is that even though people are not satisfied with their jobs, they will be less likely to quit if there are few (or no) alternatives. That is, a person would rather remain on the job than face unemployment. These dissatisfied individuals remaining on the job result in lower quit rates during times of high unemployment.

These results provide support for some of the theoretical ideas posed by Hulin et al. (1985) in a recent integrative review of the job alternatives–turnover area. Based on speculation by Michaels and Spector (1982), Hulin et al. hypothesized that the existence of job alternatives serves as a releaser of the relation between satisfaction and turnover. That is, when there are few alternatives, dissatisfied employees who wish to quit cannot do so and there will be a small correlation observed between satisfaction and turnover. When there are many alternatives, dissatisfied employees who wish to quit can, and the observed relation will become stronger. This is exactly the pattern found in the current study.

Our results suggest, however, that this explanation is incomplete. Unemployment still moderated the satisfaction–turnover relation after adjusting for quit rate. Thus, it is not just the lower turnover rate for the dissatisfied individuals that produces the effect. Rather, the causes of turnover vary as a function of unemployment. During good economic times, dissatisfaction leads employees to seek other employment, whereas satisfaction causes them to remain. During poor economic times, both the satisfied and dissatisfied individuals quit in equal numbers. People quit for reasons other than mere job satisfaction, perhaps to find better paying jobs, to return to school, or to pursue other personal interests. When jobs are plentiful, satisfaction may become more salient and more central in turnover decisions. When jobs are scarce, other considerations come into play, such as salary level, security, and future prospects.

The analysis concerning the behavioral intention component did not support the hypothesis. It was hypothesized that as the unemployment rate increased, the relation between intention to quit and turnover would remain constant. This was based on Mobley's (1977) microanalytic treatment of the behavioral-intention–turnover linkages. Our results suggest that as unemployment rate increases, the relation between intention to quit and turnover is attenuated. This is also supported by the fact that the correlation coefficients between the behavioral-intention–turnover index and national unemployment rate and the job satisfaction–turnover index and national unemployment rate were not very different from one another. If the behavioral-intention–turnover relation is moderated by economic alternatives, it indicates, contrary to Mobley's explication of the behavioral-intention–turnover linkages, that the comparison of alter-

natives may occur *after* the expressed behavioral intentions. One factor that may have led to this finding was variation across studies in the intention measure. Mobley distinguished between thoughts of quitting, which precede assessment of alternatives, which precede intention to quit. These three variables may have been mixed in the studies sampled.

The more specific occupational unemployment index did yield a stronger relation with civilian satisfaction–turnover than did the national index. However, the industrial and state rates did not appear to provide more accurate indices of employment opportunity. The reason for this difference is unclear. However, a couple of different explanations may be offered. First, the relatively lower state rate may be due to the mobility of individuals searching for employment. That is, the geographic area considered for potential jobs cuts across state lines and so the individual state rate is not as important as the occupational or national rate.

Another possible explanation is that state unemployment rates may be systematically different across good and poor economic times. For example, there may be consistently lower unemployment rates in some states and higher unemployment rates in other states, and these systematic differences are probably not related to the satisfaction–turnover relation. A similar explanation may be offered for the industrial unemployment rate and satisfaction–turnover relation. Some industries may have consistently lower or higher unemployment rates across good and poor economic times. Although these systematic differences would not be related to the satisfaction–turnover relation, they might well be related to the intention–turnover relation—thus explaining the relatively higher correlations between the state and industrial unemployment rates and the intention–turnover magnitude of relation.

As is typical of reviews cumulating studies from a diverse set of researchers, several uncontrolled sources of variance were introduced. The first was the nature of turnover itself, whether turnover was voluntary or involuntary. The majority of the sample included only voluntary turnover; however, a few studies combined both voluntary and involuntary turnover in their analyses. Many of the studies did not report whether the turnover was voluntary or involuntary, and of the studies that did give this information, most did not report the ratio of voluntary to involuntary turnover. Thus, there could have been a single case of involuntary turnover, or the sample may have been almost entirely composed of involuntary cases. It was impossible to accurately categorize the voluntary and involuntary cases, and thus impossible to separately analyze them. Second, military studies were included with nonmilitary in this analysis. When the data were reanalyzed without the military cases, there was an increase in the correlations between satisfaction–turnover and both the national and the state unemployment indices. Although the small number of military cases makes it difficult to draw conclusions, it appears that the relation between satisfaction–turnover and the national unemployment index is also strengthened when the military data are analyzed separately and adjusted for skewness. Thus, it is unclear whether the satisfaction–turnover mechanism occurs in the same manner in military cases (Hulin et al., 1985). Turnover, as assessed in the military, is a reenlistment decision made at a specified time (or time range) rather than a decision to quit the organization made

without any organizational cues (such as the end of the current enlistment period). Also, the time period of data collection tended to be longer in the military studies, possibly resulting in weaker correlations in the military sample.

Another interesting finding was the negative relation between length of turnover data collection and job satisfaction–turnover (civilian $r = -.24$, military $r = -.84$) and intention–turnover (civilian $r = -.17$, military $r = -.35$) correlations. This suggests that the intention–turnover and particularly the job satisfaction–turnover relations weaken as the time period of turnover data collection becomes longer. That is, the predictability of turnover by satisfaction and intention measures decrease with time. This effect appears to be stronger in the military sample. This may be due to the longer time periods and greater range and variance of the time periods in the military data (5 months–102 months) in comparison with the civilian data (3 months–36 months).

The results of this meta-analysis suggest that turnover researchers must consider two factors in conducting and evaluating turnover research. First, the unemployment rate at the time of a study is important because relatively little turnover will occur during periods of high unemployment and limited employment opportunity. Relations that exist during good economic times may prove difficult to replicate during bad times. Turnover researchers should include appropriate unemployment rates or the dates of their turnover data collection in their research reports. Second, the time span during which turnover data are collected should be considered. Researchers collecting data over an extremely long period of time (e.g., 3 years or more) might collect satisfaction and intention data more than once (e.g., every 2 years) to ascertain the effect of length of data collection time period on the magnitude of the relation.

References

- Atchison, T. J., & Lefferts, E. A. (1972). The prediction of turnover using Herzberg's job satisfaction techniques. *Personnel Psychology*, 25, 251–269.
- Behrend, H. (1951). *Absence under full employment* (Studies in economics and society, Monograph No. A3). Birmingham, AL: University of Birmingham.
- Crowther, J. (1957). Absence and turnover in the division of one company, 1950–1955. *Occupational Psychology*, 31, 256–269.
- Dodd, D. B., & Dodd, W. S. (1973). *Historical statistics of the United States*. Tuscaloosa: University of Alabama Press.
- Eagly, R. V. (1965). Market power as an intervening mechanism in Phillips curve analysis. *Economica*, 32, 48–64.
- Hulin, C. L., Roznowski, M., & Hachiya, D. (1985). Alternative opportunities and withdrawal decisions: Empirical and theoretical discrepancies and an integration. *Psychology Bulletin*, 97, 233–250.
- Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.
- Kraut, A. I. (1975). Predicting turnover of employees from measured job attitudes. *Organizational Behavior and Human Performance*, 13, 233–243.
- Langbein, L. I., & Lichtman, A. J. (1978). *Ecological inference*. Beverly Hills, CA: Sage.
- Ley, R. (1966). Labor turnover as a function of worker differences, work environment, and authoritarianism of foremen. *Journal of Applied Psychology*, 50, 497–500.
- Locke, E. A. (1976). The nature and consequences of job satisfaction. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 1297–1349). Chicago: Rand McNally.
- Michaels, C. E., & Spector, P. E. (1982). Causes of employee turnover: A test of the Mobley, Griffeth, Hand, and Meglino model. *Journal of Applied Psychology*, 67, 53–59.
- Mobley, W. H. (1977). Intermediate linkages in the relationship between job satisfaction and employee turnover. *Journal of Applied Psychology*, 62, 237–240.
- Mobley, W. H., Griffeth, R. W., Hand, H. H., & Meglino, B. M. (1979). Review and conceptual analysis of the employee turnover process. *Psychological Bulletin*, 86, 493–522.
- Mobley, W. H., Horner, S. O., & Hollingsworth, A. T. (1978). An evaluation of precursors of hospital employee turnover. *Journal of Applied Psychology*, 63, 408–414.
- Muchinsky, P. M., & Morrow, P. C. (1980). A multidisciplinary model of voluntary employee turnover. *Journal of Vocational Behavior*, 17, 263–290.
- Muchinsky, P. M., & Tuttle, M. L. (1979). Employee turnover: An empirical and methodological assessment. *Journal of Vocational Behavior*, 14, 43–77.
- Parasuraman, S. (1982). Predicting turnover intentions and turnover behavior: A multivariate analysis. *Journal of Vocational Behavior*, 21, 111–121.
- Porter, L. W., & Steers, R. M. (1973). Organizational, work, and personal factors in employee turnover and absenteeism. *Psychological Bulletin*, 80, 151–176.
- Price, J. L. (1977). *The study of turnover*. Ames: Iowa State University Press.
- Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage.
- Shikar, R., & Freudenberg, R. (1982). Unemployment rates as a moderator of the job dissatisfaction–turnover relation. *Human Relations*, 35, 845–856.
- Spector, P. E., & Levine, E. L. (1987). Meta-analysis for integrating study outcomes: A Monte Carlo study of its susceptibility to Type I and Type II errors. *Journal of Applied Psychology*, 72, 3–9.
- Statistical Abstract of the United States*. (Annual). Washington, DC: U.S. Government Printing Office.
- Steel, R. P., & Ovalle, N. K. (1984). A review and meta-analysis of research on the relationship between behavioral intentions and employee turnover. *Journal of Applied Psychology*, 69, 673–686.
- Waters, L. K., Roach, D., & Waters, C. W. (1976). Estimate of future tenure, satisfaction, and biographical variables as predictors of termination. *Personnel Psychology*, 29, 57–60.

(Appendixes follow on next page)

Appendix A

References of Studies Sampled

- Alley, W. E., & Gould, R. B. (1975). *Feasibility of estimating personnel turnover from survey data—A longitudinal study* (AFHRL-TR-75-54). Lackland Air Force Base, TX: Air Force Human Resources Laboratory.
- Arnold, H. J., & Feldman, D. C. (1982). A multivariate analysis of the determinants of job turnover. *Journal of Applied Psychology*, 67, 350–360.
- Butler, R. P., & Bridges, C. F. (1978). Prediction of officer retention prior to commissioning. *Journal of Occupational Psychology*, 51, 177–182.
- Coverdale, S., & Terborg, J. R. (1980). *A re-examination of the Mobley, Horner & Hollingsworth model of turnover: A useful replication* (TR 80-4). Arlington, VA: Office of Naval Research, Organizational Effectiveness Research Program.
- Dugoni, B. L., & Ilgen, D. R. (1981). Realistic job previews and the adjustment of new employees. *Academy of Management Journal*, 24, 579–591.
- Hom, P., Griffeth, R., & Sellaro, C. (1984). The validity of Mobley's (1977) model of employee turnover. *Organizational Behavior and Human Performance*, 34, 141–174.
- Hom, P. W., Katerberg, R., & Hulin, C. L. (1979). Comparative examination of three approaches to the prediction of turnover. *Journal of Applied Psychology*, 64, 280–290.
- Hom, P. W., & Hulin, C. L. (1981). A competitive test of the prediction of reenlistment by several models. *Journal of Applied Psychology*, 66, 23–39.
- Hulin, C. L. (1966). Job satisfaction and turnover in a female clerical population. *Journal of Applied Psychology*, 50, 280–285.
- Hulin, C. L. (1968). Effects of change in job satisfaction levels on employee turnover. *Journal of Applied Psychology*, 52, 122–126.
- Jackofsky, E. F., & Peters, L. H. (1983). Job turnover versus company turnover: Reassessment of the March and Simon participation hypothesis. *Journal of Applied Psychology*, 68, 490–495.
- Katzell, M. E. (1968). Expectations and dropouts in schools of nursing. *Journal of Applied Psychology*, 52, 154–157.
- Kerr, W. A. (1948). On the validity and reliability of the job satisfaction Tear Ballot. *Journal of Applied Psychology*, 32, 275–281.
- Koch, J. L., & Steers, R. M. (1978). Job attachment, satisfaction, and turnover among public sector employees. *Journal of Vocational Behavior*, 12, 119–128.
- LaRocco, J. M. (1983). Job attitudes, intentions, and turnover: An analysis of effects using latent variables. *Human Relations*, 36, 813–826.
- LaRocco, J. M., & Jones, A. P. (1980). Organizational conditions affecting withdrawal intentions and decisions as moderated by work experience. *Psychological Reports*, 46, 1223–1231.
- Lyons, T. F. (1971). Role clarity, need for clarity, satisfaction, tension, and withdrawal. *Organizational Behavior and Human Performance*, 6, 99–110.
- Michaels, C. E., & Spector, P. E. (1982). Causes of employee turnover: A test of the Mobley, Griffeth, Hand, and Meglino model. *Journal of Applied Psychology*, 67, 53–59.
- Mikes, P. S., & Hulin, C. L. (1968). Use of importance as a weighting component of job satisfaction. *Journal of Applied Psychology*, 52, 394–398.
- Miller, H. E., Katerberg, R., & Hulin, C. L. (1979). Evaluation of the Mobley, Horner, and Hollingsworth model of employee turnover. *Journal of Applied Psychology*, 63, 509–517.
- Mirvis, P. H., & Lawler, E. E., III. (1977). Measuring the financial impact of employee attitudes. *Journal of Applied Psychology*, 62, 1–8.
- Mitchel, J. O. (1981). The effect of intentions, tenure, personal, and organizational variables on managerial turnover. *Academy of Management Journal*, 24, 742–751.
- Mobley, W. H., Horner, S. O., & Hollingsworth, A. T. (1978). An evaluation of precursors of hospital employee turnover. *Journal of Applied Psychology*, 63, 408–414.
- Motowidlo, S. J. (1983). Predicting sales turnover from pay satisfaction and expectations. *Journal of Applied Psychology*, 68, 484–489.
- Motowidlo, S. J., & Lawton, G. W. (1984). Affective and cognitive factors in soldiers' reenlistment decisions. *Journal of Applied Psychology*, 69, 157–166.
- Parasuraman, S. (1982). Predicting turnover intentions and turnover behavior: A multivariate analysis. *Journal of Vocational Behavior*, 21, 111–121.
- Peters, L. H., Jackofsky, E. F., & Salter, J. R. (1981). Predicting turnover: A comparison of part-time and full-time employees. *Journal of Occupational Behavior*, 2, 89–98.
- Price, J. L., & Bluedorn, A. C. (1977). *Intent to leave as a measure of turnover*. Paper presented at the meeting of the Academy of Management, Orlando, FL.
- Price, J. L., & Mueller, C. W. (1981). A causal model of turnover for nurses. *Academy of Management Journal*, 24, 543–565.
- Rosse, J. (1983). Adaptation to work: An analysis of employee health, withdrawal, and change. *Proceedings of the Industrial Relations Research Association*.
- Spector, P. E. (1983). [Development of the Job Satisfaction Survey: A scale to measure job satisfaction in human service organizations]. Unpublished raw data.
- Spector, P. E., & Michaels, C. E. (1983). [Personality and turnover: The role of locus of control in the employee withdrawal process]. Unpublished raw data.
- Spencer, D. G., & Steers, R. M. (1981). Performance as a moderator of the job satisfaction–turnover relation. *Journal of Applied Psychology*, 66, 511–514.
- Taylor, K. E., & Weiss, D. J. (1972). Prediction of individual job termination from measured job satisfaction and biographical data. *Journal of Vocational Behavior*, 2, 123–132.
- VanZelst, R. H., & Kerr, W. A. (1953). Worker's attitudes toward merit rating. *Personnel Psychology*, 6, 159–172.
- Waters, L. K., & Roach, D. (1971). Relationship between job attitudes and two forms of withdrawal from the work situation. *Journal of Applied Psychology*, 55, 92–94.
- Waters, L. K., & Roach, D. (1973). Job attitudes as predictors of termination and absenteeism: Consistency over time and across organizational units. *Journal of Applied Psychology*, 57, 341–342.
- Waters, L. K., & Roach, D. (1979). Job satisfaction, behavioral intention, and absenteeism as predictors of turnover. *Personnel Psychology*, 32, 393–397.
- Waters, L. K., Roach, D., & Waters, C. W. (1976). Estimates of future tenure, satisfaction, and biographical variables as predictors of turnover. *Personnel Psychology*, 29, 57–60.
- Webb, W. B., & Hollander, E. P. (1956). Comparison of three morale measures. *Journal of Applied Psychology*, 40, 17–20.
- Weitz, J., & Nuckols, R. C. (1953). The validity of direct and indirect questions in measuring job satisfaction. *Personnel Psychology*, 6, 487–494.
- Youngblood, S. A., Mobley, W. H., & Meglino, B. M. (1983). A longitudinal analysis of the turnover process. *Journal of Applied Psychology*, 68, 507–516.

Appendix B

Studies Sampled in Chronological Order

Study	N'	TIME	r-st	r-it	Unemployment rates			
					NAT	IND	OCC	STA
Kerr (1948)	98	19	-.36	—	4.50	5.40	5.60	4.00
VanZelst & Kerr (1953)	340	12	-.09	—	3.30	—	—	—
Weitz & Nuckols (1953)	580	12	-.20	—	3.15	1.60	2.80	3.85
Webb & Hollander (1956) ^a	210	5	-.30	—	4.40	—	—	—
Hulin (1966)	350	6	-.28	—	4.70	—	—	—
Hulin (1968)	298	10	-.46	—	3.85	—	—	—
Katzell (1968)	1852	12	-.20	—	5.60	5.60	5.50	—
Mikes & Hulin (1968)	660	11	-.41	—	3.85	—	—	—
Lyons (1971)	156	10	-.28	+ .19	3.80	3.90	—	—
Waters & Roach (1971)	131	15	-.24	—	3.85	2.28	3.95	2.90
Taylor & Weiss (1972)	207	12	-.34	—	3.70	4.10	4.25	3.10
Taylor & Weiss (1972)	232	12	-.22	—	3.70	4.10	4.25	3.10
Waters & Roach (1973)	117	12	-.27	—	5.65	3.17	4.62	6.50
Alley & Gould (1975) ^a	5780	48	—	+ .14	4.25	—	—	—
Waters et al. (1976)	105	24	—	+ .42	5.50	3.20	4.60	5.45
Mirvis & Lawler (1977)	160	3	-.20	—	5.60	3.10	4.60	4.80
Price & Bluedorn (1977)	130	12	-.01	+ .50	7.26	6.46	7.86	3.60
Butler & Bridges (1978) ^a	465	102	—	+ .39	4.62	—	—	—
Butler & Bridges (1978) ^a	396	102	—	+ .54	4.62	—	—	—
Koch & Steers (1978)	77	9	-.14	—	7.50	3.67	5.70	9.58
Mobley et al. (1978)	203	12	-.21	+ .49	8.10	7.15	8.65	7.80
Hom et al. (1979) ^a	228	7	-.49	+ .67	7.00	—	—	6.20
Miller et al. (1979) ^a	235	6	-.38	+ .71	6.00	—	—	6.10
Miller et al. (1979) ^a	225	6	-.51	+ .66	6.00	—	—	6.10
Waters & Roach (1979)	83	12	-.25	+ .52	4.80	2.75	4.22	4.65
LaRocco & Jones (1980) ^a	260	24	—	+ .38	7.73	—	—	—
Coverdale & Terborg (1980)	65	3	-.27	+ .39	5.80	3.70	4.60	4.20
Dugoni & Ilgen (1981)	117	6	-.37	—	7.00	8.00	5.50	6.20
Hom & Hulin (1981) ^a	255	8	-.39	—	7.00	—	—	6.20
Mitchel (1981)	263	36	—	+ .29	7.51	4.20	2.87	—
Mitchel (1981)	274	36	—	+ .21	7.51	4.20	2.87	—
Peters et al. (1981)	71	18	+ .03	-.02	6.20	7.10	7.60	5.30
Price & Mueller (1981)	1091	13	-.12	+ .40	6.90	6.50	8.10	5.13
Spencer & Steers (1981)	88	12	-.13	—	5.60	5.10	5.45	8.50
Arnold & Feldman (1982)	654	12	-.24	+ .19	8.40	—	—	—
Michaels & Spector (1982)	112	7	-.20	+ .41	6.45	5.75	7.55	5.95
Parasuraman (1982)	160	12	—	+ .23	7.50	7.60	10.50	10.00
Jackofsky & Peters (1983)	265	10	-.10	—	7.40	7.80	4.50	5.25
LaRocco (1983) ^a	260	24	+ .09	+ .38	7.00	—	—	—
Motowidlo (1983)	89	20	-.17	+ .30	6.92	—	—	—
Rosse (1983)	42	13	-.07	+ .27	7.40	6.30	8.50	8.40
Spector (1983)	94	12	-.07	.00	9.70	7.60	10.60	8.20
Spector & Michaels (1983)	70	10	-.22	+ .31	8.30	6.90	9.50	7.10
Youngblood et al. (1983) ^a	1445	16	-.14	+ .13	7.70	—	—	—
Hom et al. (1984)	136	12	-.08	+ .24	7.60	6.60	8.90	8.40
Motowidlo & Lawton (1984) ^a	320	8	-.20	+ .66	5.80	—	—	—
Motowidlo & Lawton (1984) ^a	299	8	-.26	+ .61	5.80	—	—	—

Note. N' = sample size; TIME = time period in months from when satisfaction or intention data, or both, were collected, to the end of turnover data collection; r-st = satisfaction-turnover correlation; r-it = intention-turnover correlation; NAT = national; IND = industrial; OCC = occupational; STA = state.

^a Military sample.

Received September 2, 1986

Revision received December 22, 1986

Accepted December 28, 1986 ■

Examination of Avoidable and Unavoidable Turnover

Michael A. Abelson

Department of Management, Texas A&M University

Dalton, Krackhardt, and Porter (1981) suggested that examining avoidable and unavoidable turnover could improve understanding and prediction of turnover. Unavoidable leavers and stayers in the current study were found to be no different from each other, whereas both groups were significantly different from avoidable leavers on levels of satisfaction, organizational commitment, job tension, and withdrawal cognitions.

Researchers are more and more disillusioned with traditional approaches to examining the turnover phenomenon. Past reviews of the concept (i.e., Mobley, Griffeth, Hand, & Meglino, 1979), as well as multivariate studies (Arnold & Feldman, 1982; Michaels & Spector, 1982), conclude that models using individual characteristics and attitudinal variables are limited in predicting and explaining why people leave organizations. Models using these variables account for, at most, 20% of the statistical turnover variance.

One reason for this inability to better explain turnover may rest in the methodology used to examine the relationships. The traditional turnover taxonomy assumes that people leave organizations for voluntary or involuntary reasons (Bluedorn, 1978; Price, 1977). Those who leave for involuntary reasons (see Figure 1, Blocks B3 and B4) are excluded from analysis. Research used to guide turnover theory development has done so, therefore, by treating all voluntary leavers as being similar. Researchers may be willing to state that there are potential differences among those who leave voluntarily, but no empirical evidence has yet suggested that better refining the criterion turnover variable itself may increase our understanding of turnover or our ability to predict it.

Dalton, Krackhardt, and Porter (1981) presented a taxonomy that more clearly defined turnover. They suggested that a taxonomy that distinguishes avoidable (see Figure 1, Block A1) from unavoidable (see Figure 1, Block A2) voluntary turnover may improve "our understanding of the manner in which actual withdrawal decisions are made . . . [and] provide a partial explanation for the ordinarily low association between voluntary turnover and its suspected antecedents and determinants" (p. 721).

Although empirical examinations of the taxonomy suggested by Dalton et al. (1981) have not been reported, their comments suggest that those who leave for organizationally avoidable reasons are different from both stayers and those who leave for organizationally unavoidable reasons. Furthermore, this suggests

that most unavoidable leavers differ little, if any, from stayers. The definition of the turnover criterion measure itself, therefore, differs, depending on whether the traditional or expanded taxonomy is used. Although the traditional approach to examining turnover suggests that differences between stayers and all voluntary leavers are at issue, the expanded taxonomy suggests that further differentiating voluntary turnover as avoidable and unavoidable could be useful.

Hypothesized Relationships

Several hypotheses are generated to determine which turnover taxonomy is most appropriate. It is hypothesized that stayers are generally more like unavoidable leavers than avoidable leavers. The expanded taxonomy suggests that withdrawal cognitions such as thinking of quitting, intent to search, probability of finding an acceptable job elsewhere, and intent to leave are lower for stayers and unavoidable leavers than for avoidable leavers. Research that examined the relationships between withdrawal cognitions and the traditional turnover variable suggested this directionality (Arnold & Feldman, 1982; Michaels & Spector, 1982; Mobley, Horner, & Hollingsworth, 1978).

Affective responses and job-relevant perceptions for avoidable leavers are also hypothesized to be different from both stayers and unavoidable leavers. Stayers and unavoidable leavers should be more satisfied and committed and should experience less job tension than avoidable leavers (Ferris & Aranya, 1983; Mobley et al. 1979; Porter & Steers, 1973). Furthermore, significant relationships between turnover and supervisors who use their power to help employees solve work problems (Graen & Ginsberg, 1977; Sheridan, Vredenburgh, & Abelson, 1984) are hypothesized to also support the expanded taxonomy. Avoidable leavers are therefore hypothesized to experience their leader's behavior in a more negative light than do stayers or those leaving for unavoidable reasons.

Individual characteristics are the final set of variables examined in this study. The expanded taxonomy suggests that stayers and unavoidable leavers are older and more tenured than avoidable leavers. Unavoidable leavers, however, are hypothesized to be married and have more children needing care if ill than do avoidable leavers or stayers.

The author wishes to thank Thomas Bateman, John Sheridan, Stuart Youngblood, Lyle Schoenfeldt, Marietta Tretter, and two anonymous reviewers for their helpful comments on previous drafts of this article.

Correspondence concerning this article should be addressed to Michael A. Abelson, Department of Management, Texas A&M University, College Station, Texas 77843-4221.

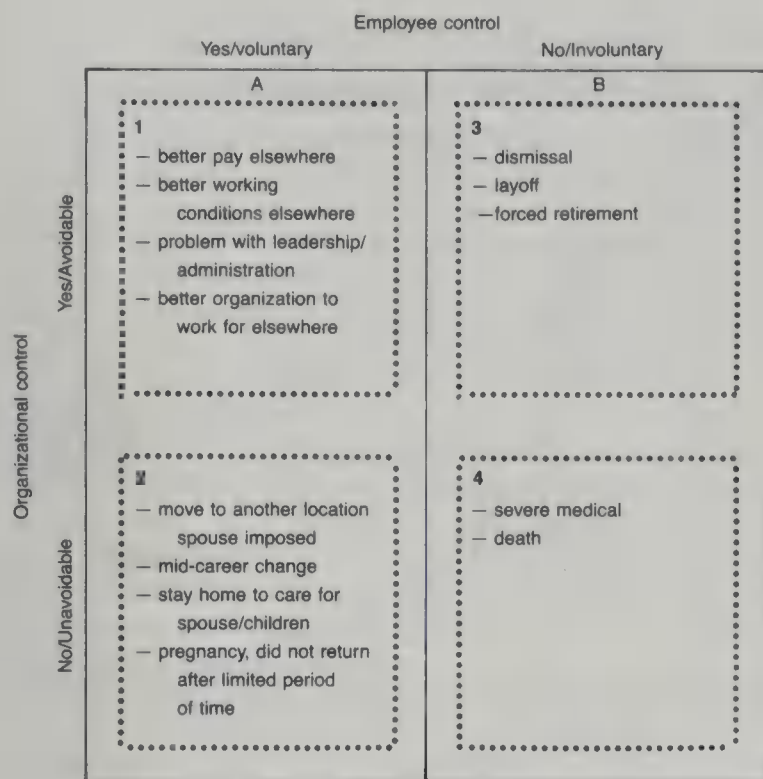


Figure 1. Expanded avoidability taxonomy.

Method

Sample

Nursing personnel from five nursing homes located in rural settings, with similar job market and economic conditions, were involved in the study. The nursing homes ranged in size from 120 to 242 beds. A total of 191 registered nurses, licensed practical nurses, and nursing aides completed questionnaires. Nursing staff members were scheduled by the director of nursing in each home to complete the questionnaire. They reported to a separate room during work hours and were allowed approximately 1 hr to complete the questionnaire. More than 95% of the staff working during questionnaire administration completed useable questionnaires. Of the 191 nursing staff in the sample, 136 remained employed and 55 left (9 left involuntarily, 30 left for avoidable reasons, and 16 left for unavoidable reasons) within 1 year of administration of the questionnaire. The annual turnover rate for all leavers was 29%. Those leaving for involuntary reasons were excluded from the analysis.

Measurement

Individual characteristics. Single-item scales were used for each of the variables. Age was in number of years. Tenure was in months functioning in their present job. Marital status was coded 0 if subjects were (currently) not married and 1 if they were (currently) married. The fourth variable was number of children school age or below who were currently living with the nurse and who would need personal care if they were ill.

Affective and job relevant perceptions. General satisfaction was measured by the Job Satisfaction Index (Brayfield & Rothe, 1951). Commitment was measured with a scale similar to that developed by Hrebiniak and Alutto (1972), and was conceptualized as an exchange in which employees compare aspects of their current job with aspects of jobs of their significant others. This approach was chosen instead of the psychological commitment approach (Mowday, Steers, & Porter, 1979) because of its conceptual similarity to the attraction-expected utility val-

ues concept hypothesized by Mobley et al. (1979) to immediately precede withdrawal cognitions. Ferris and Aranya (1983) found that neither approach was better able to predict actual turnover decisions. Job tension was measured using the Kahn, Wolfe, Quinn, Snoek, and Rosenthal (1964) 15-item scale. All three scales have frequently been used with nursing groups and are highly reliable (coincidentally, the Cronbach's alpha internal consistency reliability for each scale was .86).

Supervisory style variables were adapted from Kruse and Stogdill (1973). Leader assertiveness was a six-item scale measuring nurses' perceptions of their head nurse's use of formal power to handle disturbances and solve staff problems ($\alpha = .75$). Leader sensitivity was the other six-item scale and measured the extent the head nurse showed consideration for the nurse's feelings and maintained good interpersonal relations with staff members ($\alpha = .83$).

Withdrawal cognitions. Four single-item withdrawal cognition variables were used. These measures were identical to the one-item scales used by Mobley et al. (1978). The variables were "thinking of quitting," "intent to search," "probability of finding an acceptable job with another employer," and "intent to leave."

Turnover. Turnover was collected for 1 year and was grouped into three categories; stayers, unavoidable leavers, and avoidable leavers. Reasons for leaving were obtained from the director of nursing at each nursing home, and not from employee records. Mowday (1981) noted that there may be some attributional effects of asking nursing directors why staff left, but the nursing homes were small enough for directors of nursing services to have more accurate knowledge of why staff left than was available from company records. Furthermore, the researcher had a good rapport with the directors and was independent of the corporation, which suggested that information received from nursing directors was probably superior to corporate records even with the attributional shortcomings. In situations in which nursing directors were not sure of employees' reasons for leaving, the researcher validated reasons with nursing staff members who had previously worked with the departed staff. Examples of actual reasons for leaving used in this study are reported in Figure 1.

Determining whether reasons were avoidable or unavoidable is a complex process. Reasons given by Dalton et al. (1981), and discussions with nursing directors, were used as guidelines to determine whether turnovers were avoidable or unavoidable. The researcher did not have an opportunity to directly ask those leaving if the organization could have prevented their turnover with some reasonable action. This should be taken into consideration when examining the results.

Statistical Analysis

Means and standard deviations are calculated for the variables examined in the study. Multivariate analysis of variance (MANOVA) is performed to determine the proportion of dispersion of all of the dependent variables accounting for stayers, avoidable leavers, and unavoidable leavers. Following Spector's (1977) and others (Finn, 1974; Klecka, 1980; Tatsuoaka, 1970) suggestions, univariate F values are computed to examine the hypothesized relationships, followed by a multivariate discriminant analysis to more closely examine prediction and classification. The Behrens-Fisher method of multiple comparisons suggested by Games and Howell (1976) is used to compare means across stayers, unavoidable leavers, and avoidable leavers when the univariate F tests are significant. This method is used because it adjusts for familywise rate of Type I errors and is approximate even when cell sizes and variances are unequal.

A multiple discriminant analysis is performed to determine which of the variables best predict membership in the stayers, unavoidable leavers, and avoidable leavers groups. The direct method and not the stepwise method is used inasmuch as the stepwise method assumes theoretical justification for the order of variable entry (Tatsuoaka, 1970), and no

Table 1
Means, Standard Deviations, and Post Hoc Significance Tests for Classification Group Scores

Variable	Stayers (S)		Unavoidable leavers (UNA)		Avoidable leavers (A)		Behrens-Fisher
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Age	34.6	12.5	33.1	11.7	29.5	10.1	<i>ns</i>
Job tenure	45.1	50.6	36.2	39.3	31.3	40.2	<i>ns</i>
Marital status	0.6	0.5	0.5	0.5	0.4	0.5	<i>ns</i>
Number of children needing care if ill	1.3	1.4	1.4	1.3	1.4	1.8	<i>ns</i>
Overall satisfaction	5.1	0.9	5.1	0.9	4.5	1.1	A < UNA, A < S, S = UNA
Organizational commitment	4.5	1.4	4.3	1.3	3.8	1.3	A < UNA, A < S, S = UNA
Job tension	2.5	0.9	2.9	1.0	3.0	0.9	A > S, A = UNA, S = UNA
Leader assertiveness	5.3	0.8	5.3	0.9	5.3	0.7	<i>ns</i>
Leader sensitivity	5.3	1.3	5.5	1.0	4.8	1.3	<i>ns</i>
Thinking of quitting	2.5	1.7	2.3	1.6	4.3	1.9	A > UNA, A > S, S = UNA
Intent to search	3.2	2.1	3.4	2.3	5.3	1.9	A > UNA, A > S, S = UNA
Probability of finding an acceptable alternative job	5.4	1.5	5.2	2.0	5.4	1.4	<i>ns</i>
Intent to leave	2.4	1.5	3.0	2.1	4.8	1.9	A > UNA, A > S, S = UNA

such theoretical rationale for ordering is appropriate. A cross validation of the discriminant analysis is not performed because groups must have at least as many cases as discriminant variables (Tatsuoka, 1970), and a holdout sample used in the validation would create a violation of this assumption. (There were only a total of 16 unavoidable leavers and 13 discriminant variables.)

Results and Discussion

Table 1 shows the means and standard deviations for the variables. The one-way MANOVA was highly significant (Wilks's lambda = .634), $F(28) = 2.94, p < .001$. The analysis of variance (ANOVA) results, however, demonstrate no significant difference on mean scores between stayers, unavoidable leavers, and avoidable leavers on seven variables. This occurred with the four individual characteristic variables, both supervisory style variables, and staff's impressions of the probability of finding an acceptable alternative job. These variables were, therefore, not beneficial in determining whether the traditional or expanded taxonomy of turnover was the most useful in better understanding turnover.

Significant differences were found among the groups for each of the three perceptual and attitudinal variables and three of the four withdrawal cognition variables. As hypothesized, overall satisfaction, $F(2, 179) = 3.88, p < .05$, exchange commitment, $F(2, 179) = 3.4, p < .05$, and job tension, $F(2, 179) = 3.49, p < .05$, were different across groups. The Behrens-Fisher post hoc multiple comparisons test showed that avoidable leavers were less satisfied than those who left for unavoidable reasons ($p < .05$) and those who stayed ($p < .05$), and that stayers were no different from unavoidable leavers on levels of satisfaction. Moreover, avoidable leavers were less committed to the organization than those who left for unavoidable reasons ($p < .05$) or stayed ($p < .05$), and there were no differences between stayers and unavoidable leavers on commitment levels. Finally, job tension was significantly greater ($p < .05$) for avoidable leavers than for stayers, but no significant differences occurred between unavoidable leavers and avoidable leavers or stayers.

Thinking of quitting, $F(2, 179) = 10.6, p < .001$, intent to search, $F(2, 179) = 9.98, p < .001$, and intent to leave, $F(2, 179) = 24.5, p < .001$, were significantly different across groups. For each of these, avoidable leavers scored higher on withdrawal cognitions than did unavoidable leavers ($p < .05$) or stayers ($p < .05$). Stayers perceived withdrawal cognitions no differently than did unavoidable leavers.

An interesting trend emerged. The attitudes and withdrawal cognitions of those leaving for avoidable reasons were very different from those who left for unavoidable reasons or stayed. Avoidable leavers were less satisfied and committed and experienced greater job tension and withdrawal cognitions than did unavoidable leavers or stayers.

There were no significant differences, however, regarding individual characteristics or leader behavior perceptions. A closer examination of the means demonstrates that stayers and unavoidable leavers tended to be older than avoidable leavers, but not significantly so. Stayers and unavoidable leavers also tended to have greater tenure than did avoidable leavers. Although these relationships were not statistically significant, trends consistent with the hypothesized relationships did emerge. Further study may find relationships in the predicted direction. Family responsibility and leader behavior perceptions showed little difference across groups. The expanded taxonomy, or at least this examination of it, helps little in explaining earlier contradictory findings regarding these relationships with turnover (Mobley et al., 1979; Muchinsky & Tuttle, 1979; Waters, Roach, & Waters, 1976).

Function 1 of the discriminant analysis was highly significant and explained 84.1% of the common variance. This function was composed most significantly by three of the four withdrawal cognition variables and then by satisfaction, commitment, and job tension. A second function was not statistically significant. The analysis discriminated avoidable leavers (group centroid = -1.33) from both stayers (group centroid = .27) and unavoidable leavers (group centroid = .22). Unavoidable leavers were no different from stayers. This analysis supported the

hypothesis that the expanded taxonomy is a more appropriate turnover taxonomy.

The overall hit rate of the discriminant analysis is 80.2%; 71% of the employees actually remained. The discriminant analysis classified over 96% of the stayers accurately. Only 16% of the employees were avoidable leavers, and more than 43% of these were predicted accurately. The 9% who were unavoidable leavers were predicted accurately only 12% of the time. Most of the unavoidable leavers were inaccurately predicted to be stayers. This result further demonstrates the similarity between unavoidable leavers and stayers, giving added evidence for the expanded taxonomy.

Furthermore, these results are consistent with a postdecision justification effect. People who intend to quit and who have no external reason for explaining their quitting (avoidable leavers) may, in some cases, be modifying their attitudes and intentions to be consistent with their behavior. People who do have an external reason may not feel they have to modify their attitudes to be consistent with their behavior.

There is another possible explanation for why avoidable leavers were not better predicted (43% hit rate). Had different predictor variables been included in the analysis, such as those suggested by Mobley et al. (1979) and Steers and Mowday (1981), avoidable turnover may have been better predicted. This suggests that types of predictor variables typically examined may be as much to blame for the poor ability to predict turnover as is the lack of criterion variable precision.

Because of the inability to cross validate the results of the discriminant function classification and prediction rates, another analytic approach was used to determine the extent the avoidability taxonomy may improve prediction. A multiple regression analysis was performed using the 13 variables in the study to predict turnover. This analysis was performed using the conventional all-leavers and all-stayers approach ($n = 182$) and the avoidability taxonomy approach (excluding unavoidable leavers, $n = 165$). The adjusted R^2 for the traditional approach was .19. This is comparable with that reported by Arnold and Feldman (1982) and Michaels and Spector (1982), who both used a set of variables similar to that used in this study. When the avoidability taxonomy was used, the adjusted R^2 increased to .30. The same variables now accounted for over 50% more turnover variance just by excluding unavoidable leavers from the analysis. Another regression analysis, entering the predictor variables first and then a dummy-coded avoidable and unavoidable turnover variable, was performed. The dummy variable was highly significant, $F(1, 180) = 96.7, p < .001$, and so was the change in variance accounted for (R^2 change = .29). This latter analysis also suggests that the distinction between avoidable and unavoidable turnover is important.

Summary and Conclusion

The expanded taxonomy appears to give a more precise indication of the relation between frequently studied individual level variables and turnover than does the traditional taxonomy. First, the ANOVA and discriminant analysis demonstrated that there were no differences between stayers and unavoidable leavers for any variables studied. Second, levels of satisfaction, organizational commitment, job tension, and three of the four

withdrawal cognitions differentiated stayers and unavoidable leavers from avoidable leavers. Researchers who treat unavoidable leavers similarly to avoidable leavers may, because of methodology, be spuriously affecting their results. Practitioners using this information during in-house surveys may arrive at erroneous conclusions and take inappropriate actions. These outcomes will be affected to the extent that the firm has unavoidable turnover. Those firms with high percentages of dual-career employees, highly fluctuating economic environments (the grass is greener somewhere else in good times phenomenon), or minimally educated staff who may desire to return to school to improve their employment opportunities, may be more susceptible to spurious interpretations if the traditional and not the expanded taxonomy is used. The variables to focus on then when differentiating avoidable from unavoidable leavers seem to be the affective and withdrawal cognition variables. Age and tenure of stayers were higher, but not statistically significantly so in relation to avoidable and unavoidable leavers. There were no differences at all between the three categories for either of the family responsibility variables.

The findings presented in this article suggest that those researchers examining turnover, either in a research or an applied setting, should at least determine whether their data differentiates stayers and unavoidable leavers from unavoidable leavers on the variables of interest. If unavoidable and avoidable turnovers respond to the variables differently, there is a need to segment these different leaver categories when progressing further. Following this analytical approach will prompt us to compare our conceptual suspicions with the "reality" of the situation. More accurate theory building and application are a likely result.

References

- Arnold, H. J., & Feldman, D. C. (1982). A multivariate analysis for the determinants of job turnover. *Journal of Applied Psychology, 67*, 350-360.
- Bluedorn, A. C. (1978). A taxonomy of turnover. *Academy of Management Review, 3*, 647-651.
- Brayfield, A. H., & Rothe, H. F. (1951). An index of job satisfaction. *Journal of Applied Psychology, 35*, 307-311.
- Dalton, D. R., Krackhardt, D. M., & Porter, L. W. (1981). Functional turnover: An empirical assessment. *Journal of Applied Psychology, 66*, 716-721.
- Ferris, K. R., & Aranya, N. (1983). A comparison of two organizational commitment scales. *Personnel Psychology, 36*, 87-98.
- Finn, J. D. (1974). *A general model for multivariate analysis*. New York: Holt, Rinehart & Winston.
- Games, P. A., & Howell, J. F. (1976). Pairwise multiple comparison procedure with unequal n 's and/or variances: A Monte Carlo study. *Journal of Educational Statistics, 1*, 113-125.
- Graen, G., & Ginsberg, S. (1979). Job resignation as a function of role orientation and leader acceptance: A longitudinal investigation of organizational assimilation. *Organizational Behavior and Human Performance, 19*, 1-17.
- Hrebiniak, L. G., & Alutto, J. A. (1972). Personal and role-related factors in the development of organizational commitment. *Administrative Science Quarterly, 17*, 555-572.
- Kahn, R. L., Wolfe, D. M., Quinn, R. P., Snoek, J. D., & Rosenthal, R. A. (1964). *Organizational stress: Studies in role conflict and ambiguity*. New York: Wiley.

- Klecka, W. R. (1980). *Discriminant analysis*. Beverly Hills, CA: Sage.
- Kruse, L. C., & Stogdill (1973). *The leadership role of the nurse*. Columbus: Ohio State University Research Foundation.
- Michaels, C. E., & Spector, P. E. (1982). Causes of employee turnover: A test of the Mobley, Griffeth, Hand, & Meglino model. *Journal of Applied Psychology*, 67, 53-59.
- Mobley, W. H., Griffeth, R. W., Hand, H. H., & Meglino, B. M. (1979). Review and conceptual analysis of the employee turnover process. *Psychological Bulletin*, 86, 493-522.
- Mobley, W. H., Horner, S. O., & Hollingsworth, A. T. (1978). An evaluation of precursors of hospital employee turnover. *Journal of Applied Psychology*, 63, 408-414.
- Mowday, Richard T. (1981). Viewing turnover from the perspective of those who remain: The relationship of job attitudes to attributions of the causes of turnover. *Journal of Applied Psychology*, 66, 120-123.
- Mowday, R. T., Steers, R. M., & Porter, L. W. (1979). The measurement of organizational commitment. *Journal of Vocational Behavior*, 14, 224-247.
- Muchinsky, P. M., & Tuttle, M. L. (1979). Employee turnover: An empirical and methodological assessment. *Journal of Vocational Behavior*, 14, 43-77.
- Porter, L. W., & Steers, R. M. (1973). Organizational, work, and personal factors in employee turnover and absenteeism. *Psychological Bulletin*, 80, 151-176.
- Price, J. L. (1977). *The study of turnover*. Ames: Iowa State University Press.
- Sheridan, J. E., Vredenburgh, D. J., & Abelson, M. A. (1984). Contextual model of leadership influence in hospital units. *Academy of Management Journal*, 27, 57-78.
- Spector, P. E. (1977). What to do with significant multivariate effects in multivariate analyses of variance. *Journal of Applied Psychology*, 62, 158-163.
- Steers, R. M., & Mowday, R. T. (1981). Employee turnover and post-decision accommodation processes. In L. L. Cummings & B. M. Staw (Eds.), *Research in organizational behavior* (Vol. 3, pp. 235-281). Greenwich, CT: JAI Press.
- Tatsuoka, M. M. (1970). *Discriminant analysis: The study of group differences*. Champaign, IL: Institute for Personality and Ability Testing.
- Waters, L. K., Roach, D., & Waters, C. W. (1976). Estimate of future tenure, satisfaction, and biographical variables as predictors of termination. *Personnel Psychology*, 29, 57-60.

Received October 16, 1985

Revision received January 20, 1987

Accepted January 21, 1987 ■

Schmitt Appointed Editor, 1989-1994

The Publications and Communications Board of the American Psychological Association announces the appointment of Neal Schmitt, Michigan State University, as editor of the *Journal of Applied Psychology* for a 6-year term beginning in 1989. As of January 1, 1988, manuscripts should be directed to

Neal Schmitt
Department of Psychology
Psychology Research Building
Michigan State University
East Lansing, Michigan 48824

Manuscript submission patterns for the *Journal of Applied Psychology* make the precise date of completion of the 1988 volume uncertain. The current editor, Robert Guion, will receive and consider manuscripts until December 31, 1987. Should the 1988 volume be completed before that date, manuscripts will be redirected to Schmitt for consideration in the 1989 volume.

Application of Social Learning Theory to Employee Self-Management of Attendance

Colette A. Frayne

University of Western Ontario, London, Ontario, Canada

Gary P. Latham

University of Washington

Training in self-management was given to 20 unionized state government employees to increase their attendance at the work site. Analyses of variance revealed that compared to a control condition ($n = 20$), training in self-regulatory skills taught employees how to manage personal and social obstacles to job attendance, and it raised their perceived self-efficacy that they could exercise influence over their behavior. Consequently, employee attendance was significantly higher in the training than in the control group. The higher the perceived self-efficacy, the better the subsequent job attendance. These data were significant at the .05 level.

Kanfer's (1970, 1975, 1980) training in self-management teaches people to assess problems, to set specific hard goals in relation to those problems, to monitor ways in which the environment facilitates or hinders goal attainment, and to identify and administer reinforcers for working toward, and punishers for failing to work toward, goal attainment. In essence, this training teaches people skills in self observation, to compare their behavior with goals that they set, and to administer reinforcers and punishers to bring about and sustain goal commitment (Karoly & Kanfer, 1982). The reinforcer or punisher is made contingent on the degree to which their behavior approximates the goal. Kanfer viewed these two outcome variables in terms of informational as well as emotional feedback in order to account for cognitive as well as motoric and autonomic effects. Essentially, however, this represents a broadening rather than a change of the reinforcement contingency concept.

Training in self-regulation has been evaluated rigorously in both laboratory and clinical settings. Positive results have been obtained with regard to teaching oneself to stop smoking (Kanfer & Phillips, 1970), to overcome drug addiction (Kanfer, 1974), to reduce weight (Mahoney, Moura, & Wade, 1973), to improve study habits (Richards, 1976), and to enhance academic achievement (Glynn, 1970).

One theory that explains the effectiveness of training in self-management is social learning theory (Bandura, 1977a; 1986). This theory emphasizes the role of self-reactive influences in motivating and guiding one's behavior. The theory states that by arranging environmental contingencies, establishing specific

goals, and producing consequences for their actions, people can be taught to exercise control over their behavior.

Two social learning theory constructs that may underlie the effectiveness of training in self-management are perceived self-efficacy and outcome expectancies. Perceived self-efficacy refers to the strength of one's belief that he or she can successfully execute the behaviors required (Bandura, 1982). Such self-beliefs influence what people choose to do, how much effort they mobilize, and how long they will persevere in the face of real or perceived obstacles. For example, people who judge themselves as ineffectual in coping with environmental demands may imagine their difficulties (e.g., family obligations, transportation issues) as more formidable than they are in fact. In contrast, people who have a strong sense of self-efficacy focus their attention and effort on the demands of the situation and are spurred to an increase in effort by perceived obstacles (Bandura, 1982).

Outcome expectancies refer to beliefs concerning the extent to which one's behavior will produce favorable or unfavorable outcomes. People are prone to act on their self-percepts of efficacy when they believe that their actions will produce outcomes that are beneficial to them (Bandura, 1982). However, they are unlikely to change their behavior when they believe they can perform competently, but that the environment (e.g., supervisory or peer evaluation) will be unresponsive to their improved performance.

Few, if any, empirically based experiments have been conducted in organizational settings on the efficacy of training in self-management. Because of the diversity of clinical problems in which this training has proven effective and because of the strong theoretical rationale on which this training is based, the present study investigated its effectiveness with regard to increasing employee attendance.

Low job attendance or absenteeism is a chronic problem in organizational settings (Goodman & Atkin, 1984). Recent estimates place the annual cost of absenteeism in the United States at approximately \$30 billion (Steers & Rhodes, 1984). This is because employee absences can disrupt work schedules, increase costs, and decrease productivity.

The implicit theory underlying the present study with regard

Preparation of this article was supported in part by the Ford Motor Company Fund. The article is based on the first author's doctoral dissertation completed at the University of Washington under the supervision of the second author.

We express our gratitude to Albert Bandura, Frederick Kanfer, Edwin Locke, and Terry Mitchell for their constructive comments on earlier drafts of this article, and to H. D. Beach for his advice on conducting the training.

Correspondence concerning this article should be addressed to Gary P. Latham, Business School DJ-10, University of Washington, Seattle, Washington 98195.

to job attendance is that many people judge themselves as ineffectual in coping with environmental demands that prevent them from coming to work. Furthermore, they may believe that neither managers nor peers will change their low opinion of them even if they do increase their attendance. Support for the first assumption can be found in the revision of the Steers and Rhodes (1984) model. These authors introduced perceived inability to come to work as a critical variable affecting employee attendance. In addition, Chadwick-Jones, Nicholson, and Brown (1982) showed how cultural and normative variables can affect attendance negatively. Even a variable so straightforward as one's work schedule (flexible for most white collar jobs; inflexible for most blue collar jobs) can have a positive or negative effect on an employee's attendance. For example, people in a blue collar unionized job may find it extremely difficult to take an hour off in the middle of the workday to accompany a child to a medical doctor's office, whereas most people in white collar jobs can do this with relative ease. Thus, the former person may take the day off to accomplish what the latter person can do in 2 or 3 hr.

Support for the assumption regarding outcome expectancies is based on anecdotal evidence obtained in industry. Many people feel that once they have been labeled *poor employees* it is difficult for them to change their reputation. Support for this belief can be found in research on attitude perseverance in the face of contradictory evidence (Ross, 1977).

The primary hypothesis of the present study was that skill in self-management is a causal or independent variable that affects employee attendance positively. A second hypothesis was that the intervening variables are perceived self-efficacy and outcome expectancies. In addition to job attendance, reaction and learning measures were used as criteria for evaluating the effectiveness of the training program.

Method

Sample

A meeting was held among the first author, the director of personnel, and a representative from the union to explain how training in self-management would be a positive approach to increasing job attendance. The union agreed to support the training if certain conditions were agreed on by management, namely, that no employee would be required to receive the training, that no monetary incentive would be offered for increasing one's attendance, and that the course would be offered during normal work hours.

Employees who had used 50% or more of their sick leave received a memo from the personnel department inviting them to participate in this study. In an attempt to minimize attrition rates reported in other studies (e.g., Harris & Ream, 1978), the memo stressed that only persons who would commit themselves to eight 1-hr weekly group sessions and eight 30-min weekly one-on-one sessions should volunteer for the training. Furthermore, people who had scheduled vacation time should not enroll in the training at this time. Finally, people who were on disciplinary probation could not participate in the training.

Of the 50 individuals who were contacted, 42 volunteered to receive the training. Of these, 2 stated that they would not be able to attend all eight sessions because of scheduled vacation time. The remaining 40 individuals were randomly assigned to an experimental ($n = 20$) or a control group ($n = 20$). The people in the control group were told that they would be trained at a later date.

The mean age of the 40 employees was 44.33 years ($SD = 11.4$ years). Of the participants, 70% were men, and 30% were women. The mean number of years they had worked for the state was 7.41 ($SD = 3.14$). The individuals were employed in a maintenance department as carpenters, electricians, and painters. None of these people dropped out of the training program.

Procedure

The control group, like the experimental group, was exposed to ongoing organizational sanctions (e.g., 2 or more days off per month without a medical slip, failure to call in) regarding absenteeism. These sanctions consisted of an oral warning, a written warning, being placed on 3-months probation, and termination. The incentive for job attendance was that employees earned 8 hr sick leave each month. Hours that were not used by the end of the year would be applied to the next year. People were given compensation on retirement for the total number of sick leave hours that were not used. These policies had been in existence for 12 years. Nevertheless, the mean recorded absenteeism due to sick leave was 5.26 employee hr per week ($SD = 3.61$) in the experimental group and 4.96 employee hr in the control group ($SD = 2.16$).

The training program itself consisted of eight weekly 1-hr group sessions followed by eight 30-min one-on-one sessions. Each training group consisted of 10 people. The one-on-one sessions were conducted to tailor the training to the specific concerns of each individual and to discuss issues that the person might have been reluctant to introduce in a group setting.

The first week an orientation session was conducted to explain the principles of self-management. The second week, the reasons given by the trainees for using sick leave were listed and classified into nine categories, namely, legitimate illness, medical appointments, job stress, job boredom, difficulties with coworkers, alcohol and drug related issues, family problems, transportation difficulties, and employee rights (i.e., "sick leave belongs to me"). Of these nine categories, family problems, incompatibility with supervisor or coworkers, and transportation problems were listed most frequently. Sick leave was the focus of discussion in this session because it accounted for 49.8% of the recorded absenteeism in the organization.

The trainees were taught to develop a description of the problem behaviors (e.g., difficulty with supervisor), to identify conditions that elicited and maintained the problem behaviors, and to identify specific coping strategies. This constituted the session on self-assessment. In this session, as in all sessions, the employees were assured that their comments would not be shared with anyone outside the training group.

The third week focused on goal setting. The distal goal was to increase one's attendance within a specific time frame (e.g., 1 month/3 months). The proximal goals were the specific behaviors that the respective individual had to engage in to attain the distal goal.

The fourth week focused on the importance of self-monitoring one's behavior. Specifically, the trainees were taught (a) to record their own attendance, (b) the reason for missing a day of work, and (c) the steps that were followed to subsequently get to work. This was done through the use of charts and diaries. Emphasis was placed on the importance of daily feedback for motivational purposes as well as accuracy in recording.

In the fifth week the trainees identified reinforcers and punishers to self-administer as a result of achieving or failing to achieve the proximal goals. The training emphasized that the reinforcer must be powerful and easily self-administered (e.g., self-praise, purchasing a gift). The punisher was to be a disliked activity, easily self-administered (e.g., cleaning the garage). Each individual developed specific response-reward contingencies.

The sixth week was essentially a review of the previous six sessions. This was accomplished by asking the trainees to write a behavioral con-

tract with themselves. Thus, each trainee specified in writing the goal(s) to be achieved, the time frame for achieving the goal(s), the consequences for attaining or failing to attain the goal(s), and the behaviors necessary for attaining the goal(s).

The seventh week emphasized maintenance. Discussion focused on issues that might result in a relapse in absenteeism, planning for such situations should they occur, and developing coping strategies for dealing with these situations.

The theoretical rationale for combining these variables into one treatment package can be found in Bandura (1977b). The assumption underlying training in self-management is that the treatment package should "include as many component procedures as seem necessary to obtain, ideally, a total treatment success" (Azrin, 1977, p. 144). Empirical support for combining goal setting, feedback, and self-monitoring into a treatment package can be found in both the organizational behavior and clinical psychology literature. For example, Erez (1977) found that goal setting in the absence of feedback has no effect on behavior. Latham, Mitchell, and Dossett (1978) found that feedback in the absence of goal setting has no effect on behavior subsequent to a performance appraisal. Similarly, Simon (1979) showed that self-monitoring in the absence of goal setting has no effect on behavior. Campbell (1982) concluded that little would be gained from further attempts to tease apart the relative effects of goal setting, feedback, and reinforcers.

Results

Manipulation Checks

Goal commitment. The internal consistency of a four-item commitment measure (e.g., "To what extent will you strive to attain the goal?" "How important is it to you to at least attain the goal that was set?") administered during the final week of training was satisfactory (coefficient $\alpha = .81$). The mean of the responses to the 5-point Likert-type items was extremely high ($M = 4.73$, $SD = .22$). This restriction in range for uniformly high commitment precluded significant correlations between goal commitment and performance on the learning test, or with job attendance.

Application. A key concern was whether the trainees used the skills that were taught in the training class. Three months after the training program had been completed, the trainees were interviewed. Responses to a 5-point 13-item questionnaire on the extent to which goals were being set, feedback charts were being maintained, and reinforcers and punishers were being administered correlated significantly ($p < .05$) with job attendance ($r = .48, .45, .47$, respectively). Further corroboration that the trainees were applying the training content was obtained when the first author's visually inspected the feedback charts. Only 3 people were not keeping attendance charts systematically.

Criterion Measures

Reaction measures. Assessing employee reactions to the training was important because many trainees had argued that sick leave is a "privilege that belongs to me." Thus, it was important to determine what the trainees perceived was especially effective or ineffective about the program.

In the initial sessions, the trainees expressed hostile reactions to the training in the form of self-deprecating and aggressive comments (e.g., "I guess we are the delinquent bunch"; "the trainer is a spy for management"). A fist fight occurred in the first class as a result of name-calling between two trainees.

A 5-point five-item Likert-type questionnaire (e.g., "The training I received helped me overcome obstacles preventing me from coming to work") was completed anonymously immediately after and again 3 months after the training to measure employee reactions to the training. The coefficient alphas for this questionnaire were .70 and .73.

The employee reactions to the training were very positive immediately after training ($M = 4.32$, $SD = .55$). The employees expressed the same positive reaction 3 months after the training had taken place ($M = 4.46$, $SD = .41$). The test-retest reliability of this measure was .81.

Specifically, the trainees reported that the training enabled them to identify obstacles that prevented them from coming to work; it helped them overcome these obstacles; it led them to set specific goals for increasing job attendance; and it increased their confidence in their ability to control their own behavior.

Learning measure. It is important to understand in what way the training was effective. One criterion is learning (Kirkpatrick, 1976; Wexley & Latham, 1981). Did the trainees learn ways of responding to attendance related issues? Did they acquire problem-solving principles that enabled them to deal with coming to work effectively? To answer these questions, a learning test was developed on the basis of interviews with the supervisors of the employees.

The learning test consisted of 12 situational items with a scoring guide. A sample item is the following:

The reason I don't come to work is that I do not get along with ■ particular person with whom I work. Whenever she is on shift, I call in sick. I noted that when I do have contact with her on the job, we get into arguments. I decided to set ■ goal of "getting along with her", but it does not seem to be working. What should I do?

This methodology is based on the situational interview (Latham & Saari, 1984; Latham, Saari, Pursell, & Campion, 1980). The test was administered and scored prior to training (coefficient $\alpha = .74$) and again 3 months after the training (coefficient $\alpha = .82$) by two judges who were blind to whether the responses were from people in the training or control group. The test-retest reliability of responses to the items was .85.

An analysis of variance (ANOVA) revealed no significant difference between groups prior to training. However, the difference between groups was highly significant subsequent to training, $F(1, 38) = 6.30$, $p < .02$, $\eta^2 = .10$. The mean of the training group's performance on this test was 29.95 ($SD = 7.0$); the mean of the control group was 16.4 ($SD = 2.9$). The correlations between the premeasure and the postmeasure with job attendance were .35 ($p < .05$) and .77 ($p < .05$), respectively.

Attendance. Employee absenteeism was defined by the organization as falling into 1 of 11 categories (e.g., holiday leave, sick leave, vacation leave, jury leave, bereavement leave). Because 49.8% of the absenteeism was recorded as sick leave, this measure was used in this study. This measure was operationally defined as the number of sick leave hours taken per employee each week. Attendance was defined as the number of hours on the job when the employee was scheduled to be at work. The total number of hours the person could work on the job each week was 40. Overtime was not permitted by the organization. These two measures permitted multiple operationalism of the dependent variable of primary interest, namely, employee presence at the work site.

The test-retest reliability (stability) of the recording of sick leave assessed over a 52-week period prior to conducting the study was .38. This was in sharp contrast to the reliability of the weekly measures of attendance, namely .90. The test-retest reliability of the recording of sick leave assessed over the 12 weeks subsequent to this study was .42; the test-retest reliability of the attendance measure was .92. The correlation between the 12-week measure of sick leave and attendance was $-.64$.

A repeated measures multivariate analysis of variance based on 12 weeks of data on these two dependent variables revealed ■ significant difference between the training and the control group, $T(2, 37) = 6.67, p < .05$. Univariate F tests revealed ■ significant difference between the training group ($M = 458.4, SD = 32.7$) and the control group ($M = 403.2, SD = 22.8$) for the measure of attendance, $F(1, 38) = 5.52, p < .05, w^2 = .10$. However, the F test was only marginally significant, $F(1, 38) = 1.84, p < .10, w^2 = .027$, for the measure of sick leave ($M = 69.6, SD = 12.3; M = 83.1, SD = 18.5$, training and control group, respectively). This latter finding undoubtedly reflects the lack of stability in this absenteeism measure due to criterion contamination relative to the measure of job attendance. The accuracy of the measure of sick leave was dependent on self-report.

Intervening Variables

Of critical importance to this research was understanding why the training was effective from a psychological standpoint. Did training in self-management affect one's perceived self-efficacy and outcome expectancies? Do these variables predict job attendance?

Perceived self-efficacy and outcome expectancies were measured prior to the study, immediately after the study, and again 3 months later. The 15-item perceived self-efficacy scale followed the format used to measure self-efficacy with regard to refraining from smoking (Conditte & Lichenstein, 1981). Other items were based on comments received from supervisors ($n = 12$) and employees ($n = 10$) regarding obstacles affecting a person's coming to work. For each of the 15 items, the trainees indicated whether they felt that they would be able to come to work in each of the job attendance situations described (efficacy level) and, if yes, rated their confidence separately on a scale from 0 to 100 (efficacy strength). The coefficient alphas were .88, .91, and .89. The test-retest reliability between Time 1 and Time 3 was .92, and between Time 2 and Time 3 was .94.

Outcome expectancies (e.g., "I will not be able to meet family demands"; "I will increase my sense of accomplishment") were measured using a 15-item questionnaire that contained both positive and negative consequences, as perceived by the employee, for coming to work. These items were generated from informal interviews with supervisors and personnel officers prior to conducting the study. For each item, individuals were asked to designate on a 100-point probability scale (expressed in % units), ranging in 10-unit intervals, the probability that they would experience or achieve a particular outcome as a result of coming to work. The coefficient alphas were .67, .63, and .68, respectively. The test-retest reliability based on Time 1 and Time 3 was .74, and between Time 2 and Time 3 was .76.

There was no significant difference between the training and

control group on the measure of perceived self-efficacy or outcome expectancies prior to the training. A 2×3 repeated measures ANOVA for the self-efficacy scores revealed a significant F for groups, $F(1, 119) = 24.78, p < .05, w^2 = .24$; time, $F(2, 119) = 16.71, p < .05, w^2 = .16$; and Group \times Time interaction, $F(2, 119) = 46.02, p < .05, w^2 = .26$. The experimental group ($M = 102.3, SD = 17.5$) not only expressed higher self-efficacy than did the control group ($M = 81.1, SD = 12.4$), but this difference in perceived self-efficacy increased over time. A Pearson r between the strength of the self-efficacy measure taken immediately after training and subsequent job attendance as well as sick-leave was significant ($r = .49, r = -.40, p < .05$, respectively).

The analysis with regard to outcome expectancies did not yield any statistically significant findings. This may be because outcome expectancies were uniformly high prior to conducting the study.

Discussion

The theoretical significance of this study is that it provides an explanation of why people do or do not come to work. As Äs (1962) and Fichman (1984) have noted, absenteeism has been a social fact in need of a theory. Self-efficacy (Bandura, 1977a, 1982) is one such theory. People who come to work may be individuals who are able to overcome the personal obstacles, as well as the cultural and group norms, that were identified by Chadwick-Jones et al. (1982) as affecting one's perceived ability to come to work. People who do not come to work may be unable to cope with these influences unless their efficacy is enhanced by providing them the skill to exercise control over these variables.

Of further theoretical significance was the finding that high outcome expectancies alone will not result in employees coming to work if they judge themselves as inefficacious in overcoming personal and social obstacles to work attendance. This finding is in accord with other studies that show that low perceived self-efficacy negates the motivating potential of outcome expectancies (Barling & Abel, 1983; Godding & Glasgow, 1985; Williams & Watson, 1985).

Outcome expectancies are usually measured in terms of perceived external rewards and punishers. Social learning theory also emphasizes the role of affective self-evaluative outcomes in self-regulation through internal standards or goals. When performance falls short of the goal a person seeks to achieve, self-dissatisfaction occurs that motivates increased effort (Bandura, 1986; Bandura & Cervone, 1983). Self-motivation is regulated by both perceived self-efficacy and self-evaluation. It would be informative in future research of employees in different occupations in different industries to measure self-evaluative outcomes as well as anticipated external ones.

The practical significance of this study is fivefold. First, it showed the external validity of training in self-management for unionized workers employed by a state government. Until the present study, training in self-management had been restricted primarily to people in clinical or educational settings. Reaction, learning, self-efficacy, and job attendance measures taken 3 months after the training showed that skill in self-management brings about a relatively permanent change in cognition and affect, in addition to behavior.

Second, it showed the effectiveness of training in self-management on a dependent variable that had not been previously studied using this technique—namely, employee attendance. Employee attendance, as noted earlier in the article, has significant cost implications for organizations.

Third, employee attendance at work increased on the basis of a straightforward 12-hr training program. The concepts of goal setting and reinforcers are well-known to most trainers. What is unique to this training is the emphasis on trainees developing a contract with themselves, in addition to self-administering reinforcers and punishers to facilitate goal commitment.

Fourth, this study provided a stringent test of training in self-management. The control group, like the experimental group, was not only exposed to organizational rewards and penalties regarding attendance and absenteeism, but they had the desire to increase their attendance at work. Evidence for the latter is indicated by their attendance at the orientation session in which people were randomly assigned to the experimental or control groups. Thus, the rival hypothesis that the effects of this training were due to evaluation apprehension or attention was rejected.

Fifth, the study showed the importance of using attendance rather than a measure of absenteeism as the primary dependent variable. This point has been argued elsewhere (Latham & Frayne, 1986; Latham & Napier, 1984), but the superiority of the former measure for assessing the effects of an intervention had not been demonstrated empirically. Measures of absenteeism are typically nothing more than measures of the categorization behavior of recorders (Latham & Pursell, 1975, 1977). They typically reflect the outcome of negotiated behavior between a superior and a subordinate. That is, an absence is sometimes classified as sick leave rather than as a vacation day as a reward for good performance (Goodman & Atkin, 1984). Thus, absenteeism measures are highly contaminated (Thorndike, 1949), and their reliability is typically quite low. Had only a measure of recorded sick leave been used in this study as the sole index of absenteeism, a Type II error would have been made. The results would have shown that training in self-management had only a marginal effect on employee absenteeism.

A limitation of both the attendance and the absenteeism measures is that they ignore the distinction between voluntary and involuntary absenteeism. Some people may take sick leave because they are too ill to come to work, whereas others may have negotiated with a supervisor to record a vacation day as sick leave. These two behaviors, illness and negotiation, are very different theoretically. This lack of sensitivity in the two measures provided a highly conservative test of the training program. To overcome this problem in future studies of absenteeism, each person who stayed away from work would have to be observed, and the interobserver reliability of the observation would have to be calculated. To the authors' knowledge, this has never been done in the research literature on absenteeism.

References

- Ås, D. (1962). Absenteeism: A social fact in need of theory. *Acta Sociologica*, 6, 278–285.
- Azrin, N. H. (1977). A strategy for applied research: Learning based but outcome oriented. *American Psychologist*, 32, 140–149.
- Bandura, A. (1977a). *Social learning theory*. Englewood Cliffs, NJ: Prentice-Hall.
- Bandura, A. (1977b). Self-efficacy: Toward a unifying theory of behavior change. *Psychological Review*, 84, 191–215.
- Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, 37, 122–147.
- Bandura, A. (1986). *The social foundations of thought and action*. Englewood Cliffs, NJ: Prentice-Hall.
- Bandura, A., & Cervone, D. (1983). Self-evaluation and self-efficacy mechanisms governing the motivational effects of goal system. *Journal of Personality and Social Psychology*, 45, 1017–1028.
- Barling, J., & Abel, M. (1983). Self-efficacy beliefs and tennis performance. *Cognitive Therapy and Research*, 7, 265–272.
- Campbell, J. P. (1982, August). *I/O psychology and the enhancement of productivity*. Paper presented at the meeting of the American Psychological Association, Washington, DC.
- Chadwick-Jones, J. K., Nicholson, N., & Brown, C. (1982). *Social psychology of absenteeism*. New York: Praeger.
- Condiotte, M. M., & Lichenstein, E. C. (1981). Self-efficacy and relapse in smoking cessation programs. *Journal of Consulting and Clinical Psychology*, 49, 648–658.
- Erez, M. (1977). Feedback: A necessary condition for the goal setting-performance relationship. *Journal of Applied Psychology*, 62, 69–78.
- Fichman, M. (1984). A theoretical approach to understanding absence. In P. Goodman & R. Atkin (Eds.), *Absenteeism* (pp. 1–46). San Francisco: Jossey-Bass.
- Glynn, E. L. (1970). Classroom applications of self-determined reinforcement. *Journal of Applied Behavior Analysis*, 3, 123–132.
- Godding, P. R., & Glasgow, R. E. (1985). Self-efficacy and outcome expectations as predictors of controlled smoking status. *Cognitive Therapy and Research*, 9, 583–590.
- Goodman, P., & Atkin, R. (1984). Effects of absenteeism on individuals and organizations. In P. Goodman & R. Atkin (Eds.), *Absenteeism* (pp. 276–321). San Francisco: Jossey-Bass.
- Harris, S. N., & Ream, R. G. (1978). Follow-up strategies in the behavioral treatment of the overweight. *Behavior Research and Therapy*, 13, 167–172.
- Kanfer, F. H. (1970). Self-regulation: Research issues, and speculations. In C. Neuringer & J. L. Michael (Eds.), *Behavior modification in clinical psychology* (pp. 178–220). New York: Appleton-Century-Crofts.
- Kanfer, F. H. (1974). Self-regulation: Research issues, and speculations. In C. Neuringer & J. Michael (Eds.), *Behavior modification in clinical psychology* (pp. 178–220). New York: Appleton-Century-Crofts.
- Kanfer, F. H. (1975). Self-management methods. In F. H. Kanfer (Ed.), *Helping people change* (pp. 309–355). New York: Wiley.
- Kanfer, F. H. (1980). Self-management methods. In F. H. Kanfer & A. P. Goldstein (Eds.), *Helping people change: A textbook of methods* (2nd ed., pp. 334–389). New York: Pergamon Press.
- Kanfer, F. H., & Phillips, J. S. (1970). *Learning foundations of behavior therapy*. New York: Wiley.
- Karoly, P., & Kanfer, F. H. (1982). *Self-management and behavior change: From theory to practice*. New York: Pergamon Press.
- Kirkpatrick, D. L. (1976). Evaluation of training. In R. L. Craig (Ed.), *Training and development handbook: A guide to human resource development* (pp. 18–27). New York: McGraw-Hill.
- Latham, G. P., & Frayne, C. A. (1986, August). *The stability of job attendance*. Paper presented at the meeting of the Academy of Management, New Orleans.
- Latham, G. P., Mitchell, T. R., & Dossett, D. L. (1978). Importance of participative goal-setting and anticipated rewards on goal difficulty and job performance. *Journal of Applied Psychology*, 63, 163–171.
- Latham, G. P., & Napier, N. (1984). Practical ways to increase employee attendance. In P. Goodman & R. Atkin (Eds.), *Absenteeism* (pp. 322–359). San Francisco: Jossey-Bass.

- Latham, G. P., & Pursell, E. D. (1975). Measuring absenteeism from the opposite side of the coin. *Journal of Applied Psychology*, 60, 369-371.
- Latham, G. P., & Pursell, E. D. (1977). Measuring attendance: A reply to Ilgen. *Journal of Applied Psychology*, 62, 234-236.
- Latham, G. P., & Saari, L. M. (1984). Do people do what they say? Further studies on the situational interview. *Journal of Applied Psychology*, 69, 569-573.
- Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology*, 65, 422-427.
- Mahoney, M. J., Moura, N. G., & Wade, T. C. (1973). The relative efficacy of self-reward, self-punishment, and self-monitoring techniques for weight loss. *Journal of Consulting and Clinical Psychology*, 40, 404-407.
- Richards, C. S. (1976). When self-control fails: Selective bibliography of research on the maintenance problems in self-control treatment programs. *JSAS: Catalog of Selected Documents in Psychology*, 8, 67-68.
- Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10, pp. 174-220). New York: Academic Press.
- Simon, K. M. (1979). Self evaluative reactions: The role of personal valuation of the activity. *Journal of Cognitive Therapy and Research*, 9, 111-116.
- Steers, R. M., & Rhodes, S. R. (1984). Knowledge and speculation about absenteeism. In P. S. Goodman & R. S. Atkin (Eds.), *Absenteeism: New approaches to understanding, measuring, and managing employee absence* (pp. 229-275). San Francisco: Jossey-Bass.
- Thorndike, R. L. (1949). *Personnel selection*. New York: Wiley.
- Wexley, K. N., & Latham, G. P. (1981). *Developing and training human resources in organizations*. Glenview, IL: Scott, Foresman.
- Williams, S. L., & Watson, N. (1985). Perceived danger and perceived self-efficacy as cognitive mediators of acrophobic behavior. *Behavior Therapy*, 16, 136-146.

Received May 13, 1986

Revision received January 23, 1987

Accepted January 29, 1987 ■

Carpenter Apprentices: Comparison of Career Transitions for Men and Women

Janina C. Latack

Faculty of Management and Human Resources
Ohio State University

Bonnie L. Roach

Institute of Labor and Industrial Relations
University of Illinois

Susan L. Josephs

Faculty of Labor Education and Research
Ohio State University

Mitchell D. Levine

The Sobeco Group, Toronto, Ontario, Canada

As women are encouraged to enter nontraditional occupations, it is important to identify factors that may promote a successful transition. Data from carpenter apprentices and instructors were analyzed to compare men and women. In some aspects, based on performance and satisfaction with the apprentice program, the union, and carpentry work, the transition into carpentry is as successful for women as for men. Women give themselves a higher probability of completing the program than do men. Problem areas are evident, however. Women are employed in construction less than men, and their male co-workers have negative attitudes toward affirmative action. In terms of factors associated with success, differences between men and women emphasized co-worker acceptance, fairness in job assignments, age, and realistic expectations.

The movement of women into occupations traditionally associated with men is an important dimension of equal employment opportunity. Because women are being encouraged to enter these occupations, it is important to understand how to promote their success in these nontraditional jobs.

Although there are case study accounts of women in blue-collar jobs (Walshok, 1981), more empirical studies are needed because blue-collar women face unique difficulties in this career transition (Deaux & Ullman 1982; Terborg, Zalesny, & Tubbs, 1982.) Such studies, although few in number, are emerging. Notable examples have examined women's experiences as coal miners in Appalachia (Hammond & Mahoney, 1983), as Army ROTC members (Card & Farrell, 1983), and as steel mill workers (Deaux & Ullman, 1982).

In particular, a focus on high-paying occupations in the skilled trades is called for, because these jobs offer dramatic escape routes from the dead-end, low-paying jobs that many women have traditionally held. This study focuses on the primary route through which women enter the skilled trades: the union-employer-sponsored apprentice program. The apprentice years constitute a transitional time from school or previous employment to journeyman status as a carpenter and union member. The purpose of this study is to explore how this career transition may differ for men and women. Specifically, two is-

ssues are addressed. First, are women making a successful transition into carpentry compared with men? Second, do factors associated with success differ for men and women? At the conclusion, future research directions are recommended.

The Transition Into Carpentry

Apprentice Programs and Women

Carpenter apprentice programs last 3 to 4 years and combine classroom instruction and projects with on-the-job experience. Apprentices interact with classroom instructors, the apprentice coordinator who administers the program, and the union business agent who is responsible for obtaining and assigning jobs to apprentices and journeymen. Contractors request workers through a hiring hall, and the business agent decides which apprentices to send out on a job. Although apprentices may participate in union meetings and functions, full union membership is withheld until completion of the program, when the apprentice receives a union card and becomes a journeyman.

The apprentice program is not the only route to carpentry jobs. Individuals who obtain construction jobs can learn the trade, and the practice of purchasing union cards is relatively common. Obviously, this informal route to carpentry jobs has been largely unavailable to women. Therefore, although men have other avenues to carpentry work, for most women the apprentice program is the only way to employment in the trade.

Empirical studies of women construction apprentices have emphasized barriers to entry of women into apprentice programs (cf., Briggs, 1981; Kane, Dee, & Miller, 1977; Mapp, 1973), particularly lack of educational, psychological, and physical preparation and encouragement. A few studies have highlighted problems women face as apprentices. For example, Westley and Pinston (1982) studied federal construction projects and concluded that a key problem for women apprentices

A. Marie Sickmeier assisted with the literature review and questionnaire design. Il Jae Jung, Kye Song, and Marjorie Stassen assisted with computer analysis. Helpful comments on an earlier draft were provided by Harry Blaine, Sheila Davis, Dan Ilgen, Arnon Reichers, and Marcus Sandver.

Mitchell D. Levine is now with Hansen Consultants, Toronto, Ontario, Canada.

Correspondence concerning this article should be addressed to Janina C. Latack, Faculty of Management and Human Resources, Ohio State University, 356 Hagerty Hall, Columbus, Ohio 43210.

was gaining employment. More than two-thirds of the women surveyed reported that they could not get jobs despite advocacy of union officials. Moreover, attitudes of men were overwhelmingly negative regarding entrance of women into these occupations, and sexual harassment was pervasive.

There is some evidence that women are succeeding despite these problems. Westley and Pinston (1982) reported apprentice program completion rates of 76–100%. They also noted that employers responded positively to women's performance. Over 90% of the women apprentices in another survey reported that they found the work personally rewarding (Green, 1979).

Data on the transition of women into the skilled trades is extremely limited, but these studies suggest that women are making a successful transition despite the difficulties. To date, however, no data have directly compared problems and success for men and women during this career transition. Our research extends these previous studies by identifying a set of theory-based factors to compare the career transition experiences of men and women carpenter apprentices.

Successful Career Transitions

Outcome variables that indicate success are defined by the literature on career transitions (e.g., Latack, 1987; Louis, 1980) and socialization (e.g., Reichers, 1987; Wanous, 1980) and by features of the apprentice program. These success indicators are: satisfaction with the program, job, and union; performance; union commitment; and completion of the program.

The literature on career role transitions and newcomer socialization has underscored the importance of smoothing the transition process, arguing that initial experiences in the new role affect satisfaction, performance, commitment, and retention (Louis, 1980; Nicholson, 1984; Van Maanen & Schein, 1979; Wanous, 1980). Union leaders are concerned about the apprentices' satisfaction with the apprentice program and with the job, their performance in the classroom and on the job, and their completion of the program. In addition, the apprentice program serves as a pipeline for new members, and an important aspect of this socialization process is inculcating satisfaction with and commitment to the union organization.

Factors that may contribute to a successful transition are: expectations, anxiety, co-worker acceptance, job assignments, and acceptance by the union organization. These factors indicate whether the transition is smooth or problematic.

The literature on newcomer socialization has emphasized the importance of expectations (Wanous, 1980). Newcomers holding unrealistic expectations experience reality shock, and some studies have connected unrealistic expectations with dissatisfaction and turnover (Meglino, De Nisi, & Youngblood, 1984; see Wanous, 1980, for a review). Because previous studies of women apprentices have emphasized lack of preparation and experience, we would expect that women would be more likely than men to enter the program with unrealistic expectations.

Perhaps the single most critical variable is anxiety (Feldman & Brett, 1983; Latack, 1984; Nicholson, 1984; Van Maanen & Schein, 1979). Van Maanen and Schein argue that the most important assumption in socialization theories is that individuals undergoing any organizational transition are in an anxiety-producing situation because of loneliness, isolation, and apprehension regarding performance. In particular, women entering

nontraditional careers may experience more stress because their new situation differs along numerous dimensions from previous roles (Louis, 1980; Terborg et al., 1982). Thus, we would expect higher anxiety levels among women.

In addition, co-worker acceptance is critical during a career transition. Agents of socialization, particularly the work group, are purveyors of acceptance during the breaking-in phase and act as a primary vehicle through which socialization takes place (Wanous, Reichers, & Malik, 1984). They can speed or hinder the rate at which newcomers settle in and begin performing well (Reichers, 1984) and can ultimately provide a sense of accomplishment and competence or of failure and incompetence (Van Maanen & Schein, 1979). Previous studies have suggested that attitudes of male co-workers are negative toward women apprentices and that, instead of being accepted, women are teased and harassed. Thus, we would expect women to receive less co-worker acceptance than men receive, as reflected in the ridicule, teasing, and harassment directed toward women and in the negative attitudes of male work-group members toward the program's affirmative action efforts and toward women as carpenters.

Another critical success factor is on-the-job experience through jobs obtained and assigned by the union business agent. Previous studies suggest that women have more of a problem with this job assignment process than men. Therefore, we would expect women to view job assignments less favorably than men and to be employed less than men.

Finally, indicators of acceptance by the organization are important during career transitions. Newcomers actively search for signals that indicate acceptance in their new roles (Schein, 1978). Although official union acceptance is withheld until the completion of the program, apprentices may participate in union meetings and functions. Therefore, another critical transition variable on which men and women should be compared is the degree to which they feel accepted when they attend union meetings and functions. Because their male co-workers are unlikely to leave their negative attitudes on the job site, women may experience less acceptance than men as provisional union members.

This study also considered the differential effects of various factors on the success of men and women apprentices. The literature on organizational socialization and careers has emphasized white-collar and professional occupations. Therefore, the specification of which factors are most important to each success measure for carpentry apprentices is treated as an exploratory question. In general, however, the previous discussion suggests that the effects of realistic expectations, co-worker acceptance, job assignments, and organizational acceptance will be positive, whereas the effects of anxiety will be negative. On the basis of previous evidence concerning key problems for women, (Deaux & Ullman, 1982; Westley & Pinston, 1982), we might expect co-worker acceptance, job assignments, and organizational acceptance to be more critical in determining success for women than for men.

Method

Research Setting and Data Collection

All of the 11 district councils of the United Brotherhood of Carpenters and Joiners in a midwestern state participated in the study. A survey

was developed from background interviews conducted with apprentices, with apprentice coordinators and instructors, and from previous studies of apprentice programs (Barocci, 1973; Kane et al., 1977).

Apprentice coordinators were asked to give the questionnaires to instructors for distribution and completion by apprentices during class time. In addition, instructors were asked to evaluate each apprentice's performance in the program. Although performance ratings of job-site supervisors were solicited, because of high unemployment among apprentices (as high as 50% in some programs) and low response rate, the number returned was insufficient for analysis.

Coordinators were asked to mail questionnaires and instructor ratings directly to the researchers. After one follow-up telephone call, 8 of the 11 district councils had returned questionnaires.

Sample

Completed questionnaires were received from over 90% of the apprentices in the eight programs ($n = 406$). The sample is composed of 365 men and 41 women. Statewide, just over 8% of the apprentices were women, so our sample is slightly overrepresentative of women.

The women in the sample were significantly older than the men (mean ages were 28 and 23.7, respectively) and were more likely than the men to have some college course work. In addition, they were employed less than men in construction the year preceding the survey, and earned less money from construction work. The average woman worked approximately 40% in construction the past year, whereas the average man reported he worked over 60% of the time. Women earned an average of 25% less income from construction than men earned in the year preceding the survey.

Measures

Measures for some variables had to be developed for this study to ensure that they were applicable to carpenter apprentices. Therefore, in addition to previously developed scales, the questionnaire included two groups of items measuring various aspects of the apprenticeship transition. The first group of items dealt with the apprentice program and measured satisfaction with various aspects of the program, fairness of job assignments, and attitudes toward affirmative action efforts bringing women into the program. In addition, we wanted to assess the role of teasing and harassment. The experience of teasing and testing is common during apprenticeship and is viewed by women, in particular, as harassment, a lack of acceptance, and a hindrance to success. In fact, our interviews suggested that, for both men and women, one barometer of acceptance is that the teasing stops. The second group of items dealt with the union organization and assessed perceived acceptance at union meetings and functions as well as commitment to the union organization.

These two measures were subjected to separate factor analyses with varimax rotation. A scree test suggested that four factors should be interpreted for the apprentice program items and two factors for the union items. Items that did not load cleanly on one factor with a loading over .30 were eliminated.

The four factors for the apprentice program measures were as follows:

Program Satisfaction (17 items, $\alpha = .86$): This included satisfaction with classroom time, instructors, and overall quality of the program.

Affirmative Action Support (14 items, $\alpha = .78$): This involved attitudes toward the appropriateness of special union efforts to bring women and minorities into the program (scored so that a higher score indicated more support for these efforts).

Co-worker Acceptance (6 items, $\alpha = .81$): This measured acceptance from fellow apprentices and journeymen, feelings of fitting in, and extent of teasing and harassment experienced (scored so that a higher score indicated more acceptance and less teasing from co-workers).

Job Assignments (7 items, $\alpha = .83$): This involved views on the fairness of the job assignment process through the union business agent.

The two factors for the union measures were as follows:

Organizational Acceptance (6 items, $\alpha = .70$): This measured the extent to which apprentices felt welcome and accepted at union meetings and functions.

Union Commitment (21 items, $\alpha = .67$): This involved apprentice acceptance of union goals and his or her plans to be an active union member when finished with the program.

Additional measures assessed realism of expectations, attitudes toward women carpenters, anxiety, and the other four success indicators (job satisfaction, union satisfaction, performance, and probability of finishing).

Realistic expectations (9 items, $\alpha = .76$): This assessed the number of trade-related courses, number of years of trade-related work experience, and the extent to which the apprentice was well-informed about various aspects of the program prior to entry. Items were summed to form an indirect proxy for realism of expectations. Those who reported more courses and experience were judged to have more realistic expectations than those who did not (i.e., reporting that one was well-informed prior to entry equates to saying that expectations were in line with reality).

Attitudes toward women carpenters (21 items, $\alpha = .92$): This included measures of attitudes toward women's nontraditional role as carpenters, adapted from the Women as Managers Scale (WAMS; Peters, Terborg, & Traynor, 1974). Items from the WAMS scale were reworded to apply to the job of a carpenter.

Anxiety (20 items, $\alpha = .85$): These measures were taken from Caplan, Cobb, French, Harrison, and Pinneau (1975) and from Berkun, Bialek, Kern, and Yagi (1962). Items assess the degree to which an individual feels tense, jittery, anxious, and so forth.

Job satisfaction (4 items, $\alpha = .75$): This used Hoppock's (1935) measure. Evidence of construct validity and reliability of this scale has been presented by McNichols, Stahl, and Manley (1978).

Union satisfaction (20 items, $\alpha = .91$): These items from Hochner, Koziara, and Schmidt (1979) and from Uphoff and Dunnette (1956) assess the extent to which apprentices were satisfied with what the union provided in contract negotiation, job security, service to members, improved wages, and so forth.

Performance (2 items, $\alpha = .79$): These were instructor ratings of classroom performance using the two summary items from the Minnesota Satisfactoriness Scale (MSS; Carlson, Dawis, England, & Lofquist, 1963).

Probability of finishing: This involved self-estimated probability of completing the program, assessed with a single item, "What do you think is the probability you will finish the program?" (0 to 100%). Because few women have been in the program long enough to finish, little data on actual completion is available.

Data Analysis

Data were analyzed using the Statistical Package for the Social Sciences (SPSS; Hull & Nie, 1981; Nie, Hull, Jenkins, Steinbrenner, & Bent, 1970). In addition to *t*-tests of means differences between men and women, data were analyzed using a series of regression analyses. The attitudes toward women carpenters variable was omitted from the regression analysis for theoretical reasons. That is, we were interested in men's attitudes toward women carpenters, not in the effects of men's attitudes on men's success nor of women's attitudes on women's success. In order to test for interaction effects, interaction terms for each predictor by sex were entered into the regression equations for each dependent variable. Each interaction term was entered separately, after main effects of all independent variables had been assessed. One-tailed tests of significance were applied to coefficients for which an *a priori* hypothesis was offered.

Results

Mean comparisons for success variables and predictors are presented in Table 1. Because some of the measures were inter-

Table 1
Comparisons (t Test) for Success Variables by Sex

Variable	Women		Men		<i>t</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Age	28.0	4.8	23.7	4.2	5.30**
Education ^a	4.4	1.1	3.8	1.0	3.84**
Success measure					
Program satisfaction	60.2	11.0	62.2	10.1	1.10
Job satisfaction	20.9	2.9	20.8	2.9	.10
Union satisfaction	59.1	10.6	65.0	10.5	3.23
Performance	6.5	1.8	7.1	1.7	1.56
Probability of completion	4.8	.44	4.3	1.1	5.43*
Union commitment	17.7	3.1	16.4	3.4	2.60
Predictor					
Realistic expectations	20.4	6.1	27.6	5.8	6.86**
Anxiety	45.8	9.0	47.7	9.6	1.27
Attitudes toward women carpenters	93.6	7.9	66.5	11.8	19.05**
Co-worker support	21.0	4.5	22.4	4.4	.98
Affirmative action support	52.4	5.6	41.8	6.5	11.18**
Job assignments	26.3	7.1	29.6	6.4	2.79*
Organizational acceptance	12.5	2.5	12.3	2.6	.46

Note. *n* = 41 women; *n* = 338–365 men.
^a Values indicate that the average man is a high school graduate with some technical school, whereas the average woman is more likely to have some college education.
* *p* < .01. ** *p* < .001.

correlated, conservative alpha levels were chosen. Concerning success measures, there were no differences between men and women in satisfaction with the apprentice program, job satisfaction, union satisfaction, performance in the program, or union commitment. These results could reflect the probability of Type II error, but even at a less conservative alpha level of .05, the only other differences between men and women suggest that women may be less satisfied than men with the union, but express stronger commitment to be active in the union organization. Women rate themselves as more likely than men to complete the program, however.

In terms of predictors, significant differences between men and women confirm four of our hypotheses. First, women have less realistic expectations than men about the apprentice program. An examination of mean differences on component items of the expectations scale showed that women had fewer trade-related courses, less work experience, and rated themselves as less well-informed about all aspects of the apprentice program except pay.

Scores on attitudes toward women carpenters and affirmative action efforts reveal that not only are men more negative than women in a relative sense, but men's attitudes are also negative in an absolute sense. The mean score on these two measures indicates that the average male apprentice does not agree with *any* of the statements supporting women in this nontraditional role or supporting affirmative action efforts to bring women into the program. Women rate job assignments as less fair than do men, but the other hypotheses relative to differences in transition experiences were not supported. There are no differences between men and women on anxiety, co-worker acceptance, or organizational acceptance.

Correlational results for predictors and success measures are presented by sex in Table 2. Coefficients in the lower left portion of the table are in boldface to highlight relations between pre-

dictors and success measures. The overall pattern of the correlations confirm a priori predictions, because the sign of significant correlations for realistic expectations, co-worker acceptance, job assignments, and organizational acceptance are all positive, whereas the sign of significant coefficients for anxiety are negative. In particular, realistic expectations are positively associated with union satisfaction for both men and women. In addition, job assignments and co-worker acceptance have strong positive coefficients with several success measures. For example, realistic expectations have a positive correlation with union satisfaction for both men and women. Correlations for job assignments with satisfaction measures are positive and correlation for anxiety with satisfaction measures are negative. The overall pattern of the correlations suggests that, for both sexes, both job assignments and co-worker acceptance have the most consistent positive association with the various success measures.

The results in Table 2 do not show numerous differences between men and women, but some contrasts are noteworthy. For example, the effect of age and education on performance is negative for women and positive for men. Realistic expectations are associated with higher job satisfaction for men but are associated with higher union commitment for women. The coefficient for affirmative action support and union commitment is positive for women; the coefficient for affirmative action support with performance is negative for men. Although the correlations between co-worker acceptance and program satisfaction are positive for both samples, the association is significantly stronger for women. For men, the correlation between co-worker acceptance and probability of finishing is significantly larger than that for women. Correlational differences that were significant by Fischer *r* to *z* tests are shown in brackets. It should be noted, however, that only very large differences will be significant.

Table 2
Variable Means, Standard Deviations, and Intercorrelations by Sex

Variable	M	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Age	28 (23.7)	4.8 (4.2)	—													
2. Education	4.4 (1.57)	1.2 (1.68)	.27 (-.12)	—												
3. Realistic expectations	20.4 (27.6)	6.1 (5.8)	-.17 (-.21)*	-.26 (.01)	(.76)											
4. Anxiety	45.8 (47.7)	9.0 (9.6)	.06 (.02)	.13 (-.11)	-.04 (-.22)**	(.85)										
5. Affirmative action support	52.4 (41.8)	5.6 (6.5)	.16 (.23)**	.16 (-.08)	.13 (.00)	-.18 (-.02)	(.78)									
6. Co-worker acceptance	21.0 (22.4)	4.5 (14.4)	.12 (-.09)	.00 (.03)	.22 (.21)**	-.33* (-.24)**	-.18 (-.02)	(.81)								
7. Job assignments	26.3 (29.6)	7.1 (6.4)	-.53** [-.09]	-.52** [-.04]	.40** (.22)**	-.17 (-.10)*	-.18 (.15)**	.20 (.09)	(.83)							
8. Organizational acceptance	12.3 (12.3)	2.3 (1.7)	-.44** [-.01]	-.35* [.00]	.00 (.22)*	-.14 (-.10)*	-.08 (.11)*	.09 (.11)*	.21 (.20)*	(.70)						
9. Program satisfaction	60.2 (65.7)	10.9 (10.8)	.01 (-.02)	-.30* [.08]	.11 (.14)**	-.10 (-.15)**	-.02 (.15)**	.43** [.40**]	.27* (.39)**	(.86)						
10. Union satisfaction	59.1 (64.9)	10.5 (6.6)	-.19 (.17)**	-.30* (-.16)**	.30* (.29)**	-.36* (-.17)**	-.18 (.03)	.35* (.16)**	.55** (.60)**	.24 (.47)**	(.91)					
11. Job satisfaction	60.2 (62.2)	2.9 (2.9)	-.04 (-.06)	.12 (.04)	.08 (.30)**	-.37** (-.31)**	.10 (-.09)	.43** (.29)**	.29* (.18)	-.13 (.14)**	.25 (.30)**	.54** (.31)**	(.75)			
12. Performance	6.5 (7.1)	1.8 (1.7)	-.33* [.12]	-.18 [.20]*	-.21 (.07)	-.09 (-.05)	.01 (-.30)**	-.05 (.05)	.13 (-.06)	.12 (.06)	-.15 (-.02)	.22 (.06)	.18 (.03)	(.79)		
13. Probability of finishing	4.8 (4.3)	.44 (1.1)	.18 (-.05)	.05 (.04)	-.03 (.00)	-.09 (-.14)**	.08 (.15)**	.07 [.73]**	-.03 (.04)	-.04 (.08)	.02 (.16)**	-.04 (.07)	.24 (.22)*	.01 (.06)	—	
14. Union commitment	17.7 (16.4)	3.1 (3.4)	.16 (.13)	-.03 (.19)*	.40** (.12)	-.02 (-.03)	.38** [.01]	-.07 (.12)	.34** (.27)**	.24 (.47)**	.29* (.25)**	.09 (.36)**	.07 (.12)*	-.40* (.09)	-.05 (.11)	(.67)

Note. $n = 41$ women; $n = 338-365$ men. Internal consistency reliability estimates appear on the diagonal, with the exception of age, education, and probability of finishing (a single item). Figures for women appear first, with corresponding figures for men in parentheses. Coefficients in boldface highlight relations between predictors and success measures. Correlational differences that were significant by Fischer r to z tests are bracketed.

* $p < .05$. ** $p < .01$.

Table 3
Regression Results for Success in Apprentice Program

Independent variable	Dependent variable											
	Program satisfaction		Union satisfaction		Job satisfaction		Performance		Probability of finishing		Union commitment	
	β	F	β	F	β	F	β	F	β	F	β	F
Age	.01	.10	-.05	1.37	-.01	0.07	.07	1.8	-.02	0.08	.13	7.01**
Education	.06	1.34	-.05	1.37	.04	0.60	.09	3.25	.03	0.46		
Sex ^a	.06	.29	-.03	0.48	-.07	1.66	.26	0.11	-.18	11.59**	-.12	5.56*
Realistic expectations	.02	.68	.11	6.85*	.16	8.71**	.04	0.40	-.12	5.34*	.07	1.90
Anxiety	-.07	1.86	-.09	4.78*	-.22	20.40**	-.01	0.07	-.02	0.19	.06	1.57
Affirmative action support	.07	1.82	-.04	0.71	-.02	0.14	-.24	17.6**	-.07	1.71	-.01	0.05
Co-worker acceptance	.01	.01	.05	1.75	.10	4.19*	-.02	0.01	.46	100.89**	.07	2.30
Job assignments	.29	29.41**	.39	78.29**	.08	2.01	-.01	0.05	.02	0.08	.08	2.17
Organizational acceptance	.20	14.41**	.28	39.91**	.07	1.57	.07	1.56	.05	0.86	.40	60.52**
R^2	.21		.43		.15		.06		.24		.23	
Constant	34.20		34.93		19.58		7.30		3.24		1.92	
F	11.57**		34.58**		7.61**		3.01**		14.20**		15.19**	

Note. $n = 379$.
^a 0 = female; 1 = male.
* $p < .05$. ** $p < .01$.

In order to examine whether these relations hold when controlling for other factors, results of regression analyses are presented in Table 3. In addition to the predictor variables, age and education were included as control variables. Looking at the overall pattern of coefficients, the general hypothesis is confirmed. Although the factors that are important vary somewhat across the equations, the significant coefficients for realistic expectations, co-worker acceptance, job assignments, and organizational acceptance are positive, whereas the significant coefficients for anxiety are negative. The factors that are positively associated with more than one success measure are job assignments and organizational acceptance.

The variance explained in success measures ranges from a low of .06 for performance to a high of .43 for union satisfaction. For program satisfaction and satisfaction with the union, both job assignments and organizational acceptance have significant positive coefficients. For both union satisfaction and job satisfaction, a significant positive coefficient is observed for realistic expectations, and a significant negative coefficient is observed for anxiety. For job performance, there is a significant negative coefficient for affirmative action support. For probability of finishing, sex is significantly negative (meaning that women are more likely to finish the program than men), and co-worker acceptance is significantly positive. Finally, for union commitment, coefficients for age and organizational acceptance are each positive, and the coefficient for sex is negative.

In order to examine differences across men and women, separate interaction terms for each predictor (e.g., Realistic Expectations \times Sex) were entered into each of the full-model regression equations reported in Table 3. Interaction terms that added significantly to R^2 were as follows: For the dependent variable job satisfaction—Realistic Expectations \times Sex, $\Delta R^2 = .02$, $F(10, 395) = 7.89$, $p < .01$; Affirmative Action Support \times Sex, $R^2 = .02$, $F(10, 395) = 7.69$, $p < .01$; Job Assignments \times Sex, $R^2 = .01$, $F(10, 395) = 7.64$, $p < .01$; Organizational Acceptance \times Sex, $R^2 = .02$, $F(10, 395) = 8.05$, $p < .01$. For the dependent

variable, probability of finishing—Age \times Sex, $R^2 = .01$, $F(10, 395) = p < .05$. No other interaction terms added significantly to the variance explained. An examination of different slopes for separate regression lines showed that the positive effect for realistic expectations on job satisfaction is stronger for men than for women, that there is a positive effect for affirmative action for women but not for men, and that the positive effects for job assignments and co-worker acceptance are stronger for women than for men. For probability of finishing, older women give themselves a higher probability of finishing, whereas the effect of age is negative for men.

Discussion

The purpose of this study was to compare the relative success of men and women carpenter apprentices and examine factors associated with success. On several criteria women are as successful as men in making the career transition into carpentry. Not only are they performing as well as the men, but they are as satisfied with the apprentice program, with their jobs, and with the union as are the men. By one criteria, probability of finishing, women are more successful than men. A higher probability of finishing may indicate women's determination to succeed as they take on this nontraditional role, but it also reflects their lack of other routes to employment in carpentry.

Furthermore, concerning problems women face during this transition, the results paint a more optimistic picture than previous studies. Women report no more anxiety than men and they feel as accepted as their male colleagues at union meetings and functions. Women and men report their co-workers to be equally accepting, a finding that comes as a surprise given the negative attitudes that male colleagues hold toward women carpenters and affirmative action. These women do not report the widespread incidence of teasing and harassment that anecdotal evidence, including background interviews for this study, has suggested (Westley & Pinston, 1982). This finding is consistent,

however, with other studies reporting that, although objective indices show sex discrimination in work settings, women may not perceive it (Letvin, Quinn, & Staines, 1971). In addition, other studies measuring sexual harassment in particular have not found the expected differences between men and women (Wesman, 1984).

The transition is not problem-free, however, and our results confirm previous studies that women have difficulty obtaining the on-the-job experience crucial to completing the program and obtaining work as journeymen. Women perceive the job assignment process as significantly less fair than do men, and objective data suggest that they work less and consequently earn less than the men.

An examination of factors critical to a successful transition showed that the transition experience may be substantively different for women than for men, relative to effects of job assignments, co-worker acceptance, age, and realistic expectations. Being able to obtain work and having acceptance from co-workers, including freedom from harassment and teasing, plays a stronger role in women's job satisfaction. By contrast, having realistic information prior to entering the program is related to commitment to the union but not to job satisfaction, as was the case for men. Not surprisingly, women who view the union as taking appropriate steps to promote affirmative action are more committed to the union.

Another key difference concerns the effect of age on performance and probability of finishing. The effects for women and men are opposite: Older women receive lower performance ratings but give themselves a higher probability of finishing, whereas older men receive higher performance ratings and give themselves a lower probability of finishing. Among men, older apprentices are more likely to have had carpentry experience, thus earning them higher instructor ratings and making them less dependent on finishing the apprentice program in order to be employed.

We do not know if the negative age effect for women results from overt age discrimination on the part of instructors. However, the mean age of 28 for women compares with about 24 for men, which does not suggest that age should pose a problem with the physical and technical demands of the job. Alternatively, older women may have more difficulty with this nontraditional transition perhaps because of a longer socialization process that has portrayed male occupations as unattainable or inappropriate. These findings are particularly interesting when examined in light of other findings that were not particularly the focus of this study, namely, that older women are less likely to view the job assignment process as fair and less likely to report acceptance from co-workers. These combined findings highlight a particular subset of women that may be having problems succeeding.

Although this study represents pioneering research, the limitations of the data argue for caution in interpreting these results. First, the small sample size for women hampers the isolation of interaction effects by sex. The collection of self-report data also means that predictor-criterion relations are inflated by common method variance. Finally, the cross-sectional data do not allow us to conclude that these factors lead apprentices to be successful; rather they suggest important associations to be tested causally.

Future Research

In order to build on these findings, future research might fruitfully pursue two issues related to nontraditional career transitions for women: perceptions of harassment and age.

The women in this study did not perceive themselves as being subjected to more harassment than did men. One possible explanation to be explored in future studies is that suggested by a recent study of sexual harassment among managers (Collins & Blodgett, 1981), which found that women were expected to be able to handle whatever comes their way, especially in nontraditional fields. Thus, women carpenters may assume a *macha* (the female version of *macho*) stance, concluding that they should be able to take it and thus not perceiving the teasing and harassment.

An alternative hypothesis relates directly to the role of expectations. Realistic expectations may affect success during a transition through a "vaccination effect" (McGuire, 1964) that may generate coping strategies. If women expect a great deal of harassment, because of popular media accounts of women in the skilled trades, they may have developed coping skills to handle the harassment. They may not perceive harassment because they view themselves as coping adequately. Some support for this coping hypothesis related to realism has been found by others (Dugoni & Ilgen, 1981).

Finally, possible explanations should be sought for the negative effects of age for women. Are they, in fact, more likely to have difficulty because of their comparatively older age, or is the difficulty the result of some other variable such as education or discrimination?

If unions and contractors can capitalize on the improving economy to renew and expand affirmative action efforts for women, we can look forward to more research with larger samples of women making this nontraditional career transition. Furthermore, this study suggests that unions and contractors may be able to realistically portray an optimistic picture for women's potential success.

References

- Barocci, T. A. (1973). Apprentice Dropouts: Cause and Effect. *Manpower*, 5, 9-14.
- Berkun, M. M., Bialek, H. N., Kern, R. P., & Yagi, K. (1962). Experimental studies of psychological stress in man. *Psychological Monographs*, 76 (15, Whole No. 534).
- Briggs, N. (1981). Overcoming barriers to successful entry and retention and women in traditional male skilled blue-collar trades in Wisconsin. In V. M. Briggs and F. F. Foltman (Eds.), *Apprenticeship research: Emerging findings and future trends* (pp. 106-131). Ithaca, NY: Cornell University Press.
- Caplan, D., Cobb, S., French, R. J. P., Jr., Harrison, R. V., & Pinneau, S. R., Jr. (1975). *Job demands and worker health* (Publication No. [NIOSH] 75-160). Washington, DC: U.S. Department of Health, Education, and Welfare.
- Card, J., & Farrell, W., Jr. (1983). Non-traditional careers for women: A prototypical example. *Sex Roles*, 9, 1005-1022.
- Carlson, R. E., Dawis, R. V., England, G. W., & Lofquist, L. H. (1963). *The measurement of employment satisfaction*. Minneapolis: University of Minnesota Press.
- Collins, G. C., & Blodgett, T. B. (1981). Sexual harassment. Some see it, some won't. *Harvard Business Review*, 59, 76-95.
- Deaux, K., & Ullman, J. C. (1982). Hard-hatted women: Reflections on

- blue-collar employment. In H. J. Bernardin (Ed.), *Women in the work force* (pp. 29-47). New York: Praeger.
- Dugoni, B. L., & Ilgen, D. R. (1981). Realistic job previews and the adjustment of new employees. *Academy of Management Journal*, 24, 579-591.
- Feldman, D. C., & Brett, J. M. (1983). Coping with new jobs: A comparative study of new hires and job changers. *Academy of Management Journal*, 26, 258-272.
- Green, D. A. (1979). *Women in apprenticeship for non-traditional occupations* (Final Report). Menominee: University of Wisconsin-Stout, Center for Vocational, Technical, and Adult Education.
- Hammond, J. A., & Mahoney, C. W. (1983). Reward-cost balancing among women coal miners. *Sex Roles*, 9, 17-29.
- Hochner, A., Koziara, K., & Schmidt, S. (1979). Thinking about democracy and participation in unions. In Barbara Dennis (Ed.), *Proceedings of the 32nd Annual Meeting of the Industrial Relations Research Association*. Atlanta, GA: Industrial Relations Research Association.
- Hoppock, R. (1935). *Job satisfaction*. New York: Harper.
- Hull, C. H., & Nie, N. (1981). *SPSS update 7-9*. New York: McGraw-Hill.
- Kane, R. D., Dec, E., & Miller, J. (1977). *Problems of Women in Apprenticeship*. Arlington, VA: R. J. Associates, Inc.
- Lataack, J. C. (1984). Career transitions within organizations: An exploratory study of work, nonwork and coping strategies. *Organizational Behavior and Human Performance*, 34, 296-322.
- Letvin, R., Quinn, R. P., & Staines, G. L. (1971). Sex discrimination against the American working woman. *American Behavioral Scientist*, 15, 237-254.
- Louis, M. R. (1980). Career transitions: Varieties and commonalities. *Academy of Management Review*, 5, 329-340.
- Mapp, P. (1973). *Women in apprenticeship—Why not?* Madison: Wisconsin State Department of Industry, Labor and Human Relations.
- McGuire, W. J. (1964). Inducing resistance to persuasion. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 1, pp. 191-230). New York: Academic Press.
- McNichols, C. W., Stahl, M. J., & Manley, T. R. (1978). A validation of Hoppock's job satisfaction measure. *Academy of Management Journal*, 21, 737-741.
- Meglino, B. M., De Nisi, A. S., & Youngblood, S. A. (1984, August). *The effect of previews of job content and emotional reactions on turnover and attitudes of army trainees*. Paper presented at the 45th annual meeting of the Academy of Management, Boston.
- Nicholson, N. (1984). A theory of work role transitions. *Administrative Science Quarterly*, 29, 172-191.
- Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, E., & Bent, D. H. (1970). *Statistical Package for the Social Sciences* (2nd ed.). New York: McGraw-Hill.
- Peters, L. H., Terborg, J. R., & Traynor, J. (1974). Women as Managers Scale (WAMS): A measure of attitudes toward women in management positions. *JSAS Catalog of Selected Documents in Psychology*, 4, 27. (Ms. No. 585)
- Reichers, A. E. (1987). A interactionist perspective on newcomer socialization rates. *Academy of Management Review*, 12, 278-287.
- Schein, E. H. (1978). *Career dynamics: Matching individual and organizational needs*. Reading, MA: Addison-Wesley.
- Terborg, J. R., Zalesny, M. D., & Tubbs, M. E. (1982). Socialization experiences of women and men graduate students in male sex-typed career fields. In H. J. Bernardin (Ed.), *Women in the work force* (pp. 124-155). New York: Praeger.
- Uphoff, W. H., & Dunnette, M. D. (1956). *Understanding the union member*. Minneapolis: University of Minnesota Industrial Relations Center.
- Van Maanen, J., & Schein, E. H. (1979). Toward a theory of organizational socialization. In B. M. Staw & L. L. Cummings (Eds.), *Research in Organizational Behavior* (Vol. 1, pp. 209-264). Greenwich, CT: JAI Press.
- Walshok, M. L. (1981). *Blue collar women: Pioneers on the male frontier*. Garden City, NJ: Anchor Press.
- Wanous, J. P. (1980). *Organizational Entry*. Reading, MA: Addison-Wesley.
- Wanous, J. P., Reichers, A. E., & Malik, S. D. (1984). Organizational socialization and group development: Towards an integrative perspective. *Academy of Management Review*, 9, 670-683.
- Wesman, E. C. (1984, August). *Development of a perceived sexual harassment scale*. Paper presented at the 45th annual meeting of the Academy of Management, Boston.
- Westley, L. A., & Pinston, G., Jr. (1982). *A territorial issue: A study of women in the construction trades*. Washington DC: Wider Opportunities for Women and the Center of National Policy Review.

Received October 2, 1985

Revision received January 30, 1986

Accepted November 21, 1986 ■

Types and Choices of Performance Feedback

Daniel R. Ilgen
Michigan State University

Carol F. Moore
Allstate Research and Planning Center, Menlo Park, California

Although research has clearly demonstrated that specific and timely feedback to individuals is beneficial to task performance, little attention has been paid to the content of the feedback on the most typical type of work tasks—tasks in which high performance along both quality and quantity dimensions is desired and in which quality and quantity are inversely related at high levels of performance. In a two-phased study, this research placed one group of 132 subjects on a task that required both quality and quantity performance. Performance feedback was then presented on either quality or quantity, or on both dimensions. In a second set of conditions the same types of feedback were made available to 90 subjects, but the subjects were now free to access or not to access the feedback. The results showed that the nature of the feedback affected performance on both dimensions and that allowing feedback choice improved overall performance. Results are discussed in terms of the benefits of feedback choices on multifaceted performance tasks.

Both theory and practice strongly advocate the availability of specific and timely feedback for enhancing task performance (Ammons, 1956; Ashford & Cummings, 1983; Ilgen, Fisher, & Taylor, 1979; Larson, 1984; Locke, Shaw, Saari, & Latham, 1981). There have been numerous demonstrations, both in the laboratory and the field, that feedback enhances performance on single criterion tasks when it is made specific to the performance criterion. For example, Locke et al. (1981) reviewed a number of laboratory and field studies in which performance improved when goals were set and feedback was provided. Becker (1977) demonstrated that energy consumption could be reduced by presenting residents with daily feedback about the number of kilowatts used, and Komaki, Barwick, and Scott (1978) found that the number of safe behaviors displayed by bakery employees increased when feedback was provided.

For the most part, feedback effects on performance have been demonstrated for tasks in which the primary performance criterion was quantity (i.e., number of units produced for a specified time period). Yet, most tasks encountered at work have both a quantity and a quality dimension. Furthermore, at their upper extremes, the two dimensions become interdependent. Maximizing quantity can only be done at the expense of some quality and vice versa. When both quality and quantity are important performance criteria and are also inversely dependent, what constitutes "specific and timely" feedback is uncertain. The problem is made even more complex because the very act of presenting feedback to the performer usually means that that

person must take time away from the task to study the feedback, particularly on complex tasks. Thus, receiving feedback often interferes with the extent to which quantity goals can be reached when the time needed to receive and process feedback reduces the time available to work on the task.

Feedback is most valuable for improving performance when it provides information useful for choosing specific performance-enhancing behaviors (Ashford & Cummings, 1983). When both high-quality and high-quantity performance is desirable, feedback should provide information that allows the recipient to improve the quality or to vary the speed or intensity of work on the task. The exact effect of such information depends on the specific task and the nature of the task's quality and quantity interdependence. It also depends on whether feedback functions to direct behavior or to motivate it (Locke, Cartledge, & Koeppel, 1968; Payne & Hauty, 1955). In the former case, the function is primarily one of learning the correct response. In the latter, it is one of stimulating action. Although current research on performance feedback tends to focus on the latter function, whereas earlier work tended to focus on knowledge of results (e.g., Ammons, 1956), the two functions rarely can be isolated. The research reported here will present feedback that stresses information (the correct spelling of words) and motivation (speed of task completion).

Allowing performers to choose feedback may resolve some of the conflict between the need for feedback and its cost in terms of time. If people access feedback only when they need additional information, some of the redundancy created by giving feedback to everyone could be reduced. This conclusion is based on the assumption that people will seek feedback about their own performance only when such feedback provides information useful for improving performance on a task on which they desire to perform well. It also assumes that they will not seek feedback when it duplicates information that they already possess. These assumptions are consistent with the findings of Ashford and Cummings (1983), who viewed feedback as a personal resource that individuals actively monitor and then select when the feedback is perceived to be valuable for reaching per-

This research was supported, in part, by Grant #N00014-83-K-0756 from the Office of Naval Research, Organizational Effectiveness Unit. This support is greatly appreciated; we wish to note, however, that the ideas expressed here are those of the authors and are not necessarily endorsed by the supporting agency.

We also want to thank Cheri Ostroff for her help on some of the analyses.

Correspondence concerning this article should be addressed to Daniel R. Ilgen, Department of Psychology, Michigan State University, East Lansing, Michigan 48824-1117.

formance goals. Their views are also consistent with recent work on the importance of feedback for the individual in making internal cognitive comparisons with his or her own performance standards (Bandura & Cervone, 1983; Campion & Lord, 1982; Earley, 1986; Matsui, Okado, & Kakuyama, 1982). In all these cases, the data imply that performance feedback provides a basis for making judgments about the extent to which some goal or standard is (or is not) being accomplished and that the hedonic consequences of these judgments often affect the amount of effort devoted to the task.

The Present Research

This research was designed to address several feedback issues on a task that demanded feedback functioning both to direct behavior and to motivate it. Specifically, both high-quality performance and speed were important to accomplish successfully the task goal. Participants were either presented with feedback (knowledge of results on a spelling task) or allowed to choose various forms of feedback. The design allowed for exploring the effect of each type of feedback and the effect of feedback choice on performance. It was also possible to explore the correlates of feedback-seeking behavior. In particular, it was predicted that, assuming a constant desire for good performance, those who believed that the feedback had higher information value would seek more feedback than those who perceived the feedback to be of little value.

Method

Participants

Two hundred twenty-two undergraduate students enrolled in an introductory psychology class participated in the study and received class credit for their participation.

Design

A 3×2 design (three levels of feedback crossed with two levels of choice) with an additional control group was used. The three feedback conditions were quality feedback, quantity feedback (expressed in terms of the amount of time taken on the task), and both quality and quantity feedback. Each of these will be described in detail later. For the three feedback conditions, two forms of control over feedback were used. In one case, participants were given feedback at the end of each trial. In the other condition, they received feedback only if they chose to receive it. Finally, for the control condition, participants received no performance feedback. Obviously, without any feedback, control versus no control of feedback was meaningless; therefore, the control group was not included in the 3×2 design.

Task

The task involved proofreading nine paragraphs for spelling errors. The paragraphs averaged about 75 words in length and had, on the average, 5 misspelled words per paragraph.

Paragraphs were presented to the subjects on a video display monitor (a 12 in., green phosphorous ADM Information Display) connected to a microcomputer system (Cromemco Z-2D System 2). A paragraph was presented on the upper portion of the video screen with numerals printed on the far left of the display to identify lines in the paragraph. In order to respond to an error in the paragraph displayed, the participant

entered the line number in which the error appeared on the keyboard in front of him or her. After the number, the person entered the word as he or she thought it should be correctly spelled. This was followed by pressing the *return* key.

A question mark appeared on the screen immediately following the *return* response. If the person wished to enter another error, the process just described was repeated. When the person could identify no more spelling errors, *no* was entered in response to the question mark.

At this point, the feedback condition was administered. Following the feedback manipulation, the person typed the numeral 1 followed by *return*, and the next paragraph appeared on the screen.

All participants, regardless of the feedback condition, were assigned the same performance goals. These goals were described in terms of both the number of misspellings to identify correctly (quality) and time/speed (quantity). Specifically, all were asked to achieve a final score of 35 in 22.5 min. They were told that their score was the number of correct identifications of misspellings minus .25 times the number of incorrectly identified words. (This formula was a correction for guessing.) The subjects were also told that they were required to average about 2½ min per paragraph in order to accomplish their time goal. The final performance goal and the time limit were based on pilot work that showed the combination of these two to be difficult but attainable.

Experimental Conditions

Recall that three performance feedback conditions were created—quality only, quantity only, both quality and quantity—as well as a control group that received no feedback. Each form of feedback will be described now as it existed when the feedback was presented on every trial. Slight modifications under the choice manipulation will be described later.

For the quality-only condition, two sets of information (feedback) about the spelling of words were presented. The first appeared on the screen immediately after the participant identified ■ word he or she believed to be an error by typing in his or her spelling of that word. If an incorrectly spelled word was selected and then spelled correctly, the word *correct* appeared on the screen. If either ■ correctly spelled word was selected as an error or the correct spelling was not given for an incorrectly spelled word selected, the word *incorrect* appeared on the screen. In addition, once the person indicated that no more incorrectly spelled words could be located in the paragraph, the following information appeared on the screen: (a) the correct spelling for all incorrectly spelled words in the paragraph, (b) the total number of correct and incorrect words the person had identified on all paragraphs up to that point (termed *running total* on the screen), and (c) the person's performance score up to that point.

For quantity-only feedback, when the participants indicated that they were finished with the paragraph, the following information appeared on the screen: (a) the amount of time spent working on the paragraph, and (b) the average amount of time spent on each paragraph up to that time.¹ Subjects in the quality-and-quantity feedback condition received the information about quality and about time as described above. In the no-feedback condition, participants were simply presented with a new

¹ Note that time is, in our opinion, the critical issue when quantity is the focus of evaluation on any task, because quantity is meaningful only in units per some specified period of time. However, proofreading, when defined in terms of identifying errors, does not have a direct quantity dimension. Although number of errors identified is quantity, misidentification of errors is not. Given the nature of the goal, measuring time on task more closely simulated the behavioral implications of quantity on this task than did the number of words identified. Therefore, time feedback was provided as indicated.

Table 1
Means, Standard Deviations, and Intercorrelations of Individual Differences and Responses

Variable	M	SD	N	1 ^a	2	3	4	5	6	7	8
1. Spelling ability	19.06	3.89	222	—							
2. Self-esteem	32.78	3.82	221 ^b	.09	—						
3. Performance	27.63	7.36	222	.69**	.11	—					
4. Amount learned during task	1.29	1.06	222	.06	.00	.17**	—				
5. Amount learned posttask	2.02	1.27	222	.06	.08	-.03	.68**	—			
6. Number of words chosen	35.95	6.39	222	.39**	.07	.72**	.11	-.14*	—		
7. Number of trials feedback chosen	5.96	3.20	90	.09	-.07	.24*	.35**	.20	.27**	—	
8. Time	22.05	6.08	222	-.32**	-.14*	-.11	-.17**	-.02	.15*	.18	—

^a The number of cases for each correlation is the smaller of the two *N*s. ^b One person did not take the self-esteem scale but completed all of the other measures.

* $p \leq .05$, two-tailed. ** $p \leq .01$, two-tailed.

paragraph when they indicated that no further errors could be found in the just-completed paragraph.

One hundred thirty-two people (30 per condition) received one of the four conditions just described. An additional 90 people (30 per feedback condition, excluding no feedback) were given a choice of whether they wanted to receive feedback at the end of each trial. In the choice condition, when subjects indicated that they had completed the paragraph, a message on the screen asked them to choose one of two options and instructed them about which key to use for selecting each option. The options were to continue on to the next paragraph or to receive feedback on past performance. The nature of the feedback, if chosen, depended on the group to which the person was randomly assigned.

Procedure

All subjects were run individually. When participants reported to the experiment, they were taken to the experimental room. This room had two attached 6' × 7' cubicles. One cubicle was set up with a table and chair for completing paper and pencil instruments and the other with the microcomputer and monitor.

After a brief introduction to the study, the person entered the cubicle with the table and chair and completed a spelling test. Upon its completion, the person was taken to the cubicle with the computer, where the task was explained in detail, the goal was presented, and a practice proofreading task with feedback (which matched the person's feedback condition) was administered. Following the practice session and the answering of the participant's questions, the experimental trials began. Once the nine trials were completed, the person returned to the original cubicle to complete a posttask questionnaire and to be debriefed.

Measures

Ability. A list of 26 commonly misspelled words, which included many misspellings, was presented to the participants with instructions to correct the errors. The words were randomly selected from the book *335 Real Spelling Demons for College Students* (Furness & Boyd, 1959).

Self-esteem. Included in the pretask questionnaire, along with the spelling ability test, were a number of background information items, belief and attitude scales, and the Rosenberg (1965) Self-Esteem Scale. Only the latter was of interest to us in this context. This frequently used scale contains 10 items and typically has reported internal consistency reliabilities in the .80s.

Behaviors. Six measures of behavior were obtained. The first was an index of *performance*, which was constructed exactly as described to the

subjects (i.e., the number of properly corrected misspelled words minus .25 times the number incorrect).

Two behavioral measures tapped the amount that the person learned over the course of the experiment. The first of these was labeled *amount learned during task*. Among the words incorrectly spelled, nine were misspelled in two different paragraphs. When a participant either correctly identified one of these words and then misspelled it or failed to identify it as misspelled the first time it appeared, we assumed that that person had the potential for learning how to spell it by the second appearance of the word. Therefore, if either of these conditions occurred, and the person spelled the word correctly on the second encounter, one point was added to the person's score on the amount-learned-during-task variable. On the other hand, if the person got the word correct the first time it appeared, we did not count performance on its second appearance, because we felt that spelling the second time was, more than likely, due to knowing how to spell the word all along rather than to having learned it during the task.

A second learning variable, *amount learned posttask*, was based on the person's responses to the same nine words imbedded in a 42-item spelling test that was included in the posttask questionnaire. In this case, nearly the same decision rule, based on performance when the word appeared in the task, was used to decide how to score correct spellings on the paper-and-pencil test. In this decision, however, we included misses or misspellings on either the first or second appearance of the word. The second learning measure was included primarily to detect learning that took place after the second occurrence of a word. In this case, a word incorrectly spelled both times it appeared in text may have been learned from the feedback received the second time. The first learning measure would not detect this. The second measure also detected conditions in which words spelled incorrectly on their first appearance led to a correct guess on the second appearance, but not when the word was presented after completing the task. The correlation between the two learning measures was .68.

A behavioral measure was also constructed to assess the frequency with which the person chose words as misspelled (termed *number of words chosen*). This index was simply the number of words the person selected summed over all nine paragraphs.

Another behavioral measure was the total *time* the person spent on the task. Time was measured beginning with the first presentation of the first paragraph and ending when the person indicated that he or she had completed the ninth paragraph.

Finally, for those who were allowed to choose whether to receive feedback, the number of times it was chosen was recorded. This variable was labeled *number of times feedback chosen*.

Table 2
Cell Means for Comparisons of Feedback Conditions on Behavioral Measures

Variable	Feedback condition			<i>F</i>	<i>p</i>
	Quantity	Quality	Both		
Performance	24.54	28.83	29.86	9.85	≤.001
Amount learned during task	0.95	1.73	1.59	10.64	≤.001
Amount learned posttask	1.86	2.41	2.24	3.19	≤.05
Number of words chosen	32.87	37.29	37.44	12.28	≤.001
Time	19.33	24.97	22.28	14.80	≤.001

Results

In order to ensure that the feedback groups did not differ in terms of the two individual difference measures—spelling ability and self-esteem—one way analyses of variance (ANOVAS) were run within choice conditions and across feedback levels in order to include the no-feedback group in the design. For the no-choice condition there were four feedback levels (quantity, quality, both, and no feedback), and for the choice condition there were three (quantity, quality, and both). No significant differences in individual difference measures across experimental conditions were found.

Table 1 presents the means and standard deviations of the individual difference and the behavior measures as well as their intercorrelations. First, it should be noted that ability correlated quite highly with performance on the task. Those with higher ability also correctly identified more misspelled words and completed the task more quickly than those with lower spelling ability. With respect to performance, those who performed better on the task also scored higher on the index of learning during the task, although the magnitude of this correlation was not very high. The better performance of those who identified more words and of those who chose feedback more frequently was of more interest.

Cell means by conditions are presented in Tables 2 and 3. Because there were no significant interaction effects, only the means for simple main effects are presented. Also, performance in the no-feedback conditions did not differ significantly from that in the quantity-only feedback condition and, for that reason, was not included in further analyses.

For all these performance variables, performance was superior when spelling information was available (see the first three rows of Table 2). Comparisons among group means (Winer, 1962) showed that the quality and quality-and-quantity feedback conditions differed from each other only in the case of the amount learned during the task. In this case, those who received only quality feedback performed better than those who received both quality and time information. Interestingly, with respect to time, quality feedback alone took longest (as expected) but, if time information was provided as well, participants were able to maintain overall performance similar to the quality-only group yet accomplished this in significantly less time than those who received only quality feedback (22.28 vs. 24.97 min.).

Table 3
Cell Means for Comparisons of Choice Effects on Behavioral Measures

Variable	Choice		<i>F</i>	<i>p</i>
	No	Yes		
Performance	27.91	27.56	0.12	<i>ns</i>
Amount learned during task	1.38	1.47	0.32	<i>ns</i>
Amount learned posttask	2.12	2.22	0.31	<i>ns</i>
Number of words chosen	36.82	34.82	5.44	≤.05
Time	23.63	20.61	12.68	≤.001

A final set of analyses was conducted only with persons who were allowed to choose feedback. In this case, feedback-choice behavior was correlated with individual difference and behavioral measures, both within feedback condition and across the three conditions. These data are reported in Table 4.

First, with respect to individual differences, only spelling ability predicted choice behavior and then only in one condition. When the feedback was given about how to spell words, those with higher levels of spelling ability chose to receive the feedback less often.

For the overall sample, the number of trials after which feedback was chosen correlated positively with two measures of performance: overall performance ($r = .24, p < .05$) and the amount learned during the task ($r = .35, p < .01$). However, this effect occurred only for the persons who received both time and quality feedback. In fact, for this group the amount learned posttask also correlated significantly with the amount of feedback chosen as did the number of words the person selected as misspelled. Finally, choosing quality feedback slowed down

Table 4
Correlations Between Amount of Feedback Chosen and Both Individual Difference and Performance Measures

Variable	Type of feedback chosen			
	Overall (N = 90)	Quantity (N = 30)	Quality (N = 30)	Both (N = 30)
Spelling ability	.09	.23	-.45**	.20
Self-esteem	-.07	.03	-.30	.05
Performance	.24*	.29	-.29	.39*
Amount learned during task	.35**	.32	.27	.36 ^a
Amount learned posttask	.20	.03	.00	.43*
Number of words chosen	.27**	.23	-.08	.40*
Time	.18	-.24	.48**	.04

^a $p < .052$, two-tailed.
* $p < .05$, two-tailed. ** $p < .01$, two-tailed.

performance on the task ($r = .48, p < .01$), but only if no information was given about performance with respect to time. When time information was given along with quality feedback, the correlation between time and the number of times feedback was chosen dropped to zero ($r = .04$). Thus, choosing feedback appeared to have a positive effect on performance but did not significantly increase the amount of time necessary to complete the task in which performing quickly was a goal given to the participants.

Discussion

Two issues were of interest in this research. First, we wanted to address the effects of different types of feedback on task performance when performance was evaluated both in terms of quality and the time taken to perform the task. Although a large number of tasks confronted at work are judged on both these criteria, rarely has research on feedback studied tasks on which both time and quality were important and interdependent. Second, feedback-choice behavior was investigated both in terms of antecedents and consequences.

Ignoring feedback choice for a moment, the results supported the general conclusion that feedback information had the greatest effect on those task behaviors that most closely matched the feedback. In general, those who received feedback about the speed of their performance completed the task more quickly than those who did not receive such information, and those that received feedback about the quality of their performance scored higher on quality-related measures (see Table 2).

Although these results are encouraging from the standpoint of the value of feedback, they are not particularly interesting. Of greater interest was the joint effect of quality and quantity feedback when task goals stressed both. Here the data imply that providing quality information slowed down task completion, but, if the participants also received information about their speed, they responded by working more rapidly on the task while still having to take time to receive the qualitative feedback. Those who received both time and quality information completed the task, on the average, 2.5 min quicker than those who were only given quality information (see Table 2). This represented a 10% reduction in time needed for task completion. It is important to note that the increased speed was not purchased in terms of lower performance on any of the quality measures; for all three quality measures there were no significant differences between groups that received both quality and time information and those that received information only on quality. On the other hand, those who only received information about the speed of their task performance, performed worse than the other two groups on the quality measures. The results imply that, when tasks demand quality performance under time pressures, feedback should be available on both dimensions in order to influence both time and quality performance.

The results related to the option to choose feedback suggest that the inherent conflict between the need for information about performance quality and the time it takes to receive and process the information can be alleviated somewhat. Those allowed to choose feedback completed the task significantly more quickly than those not allowed to do so, yet the performance of

the two groups did not differ significantly on any of the other performance variables (see Table 3). Although it is conceivable that freedom of choice alone sufficiently motivated subjects to speed up performance in ways other than simply choosing less feedback, post hoc analyses indicated that this was unlikely. Because 36 of the 90 participants who were given the option to choose or not choose feedback chose it on all 9 trials, the time required for these participants was compared with the time of those who were required to receive feedback. Using a 2×3 ANOVA with time as a dependent variable, no effect was found for having or for not having the option to choose. These data implied that the number of trials on which feedback was chosen rather than reacting more quickly to chosen feedback in the choice condition led to the time differences.

A closer look at the choice behavior itself showed that those who chose feedback more frequently performed better on two of the three quality measures overall, but that this relationship was statistically significant only for those who received both quality and time feedback. Given the time-indexed goal, it is important to note that, although higher performance correlated with more frequent feedback choice, those who chose feedback more frequently when it contained both sets of information did not take more time to complete the task. In contrast, those who received only quality feedback took longer to complete the task (see Table 4).

The data in Table 3 are consistent with the integration of Ashford and Cummings' (1983) view of feedback as a resource of the individual and the psychological mechanisms described by social learning theory (Bandura & Cervone, 1983) and control theory (Campion & Lord, 1982; Taylor, Fisher, & Ilgen, 1984). Both of the latter theories assume that feedback provides information to the performer that, when compared by the performer to his or her performance standard, creates some affective reaction that has motivational properties. In response to the affect created, the theories assume that the individual chooses whether to invest further effort in working on the task or attempts to change his or her standard for performance.

In this study, the performance standard was the goal expressed in terms of both time (quantity) and proofreading performance (quality). The participants in the research needed two types of information to perform well on the task. One type was the speed at which they were performing the task and the other was feedback about the correct spelling of words. Given the nature of the task of spelling, good spellers should need feedback on their performance less than poor spellers because they have internal standards against which to compare their work. If good spellers recognize this, they should choose feedback less frequently than poor spellers on a task for which time is important and feedback takes time. Table 3 shows that spelling ability did correlate negatively with the choice of feedback but only for the quality feedback condition ($r = -.45$, between frequency of choice and spelling ability). In the other two conditions, requesting feedback also provided information about time. Therefore, we would expect that feedback choice would be driven less by the person's ability to spell in those conditions than was the case in the quality-only condition. The lack of a significant correlation between choosing feedback and spelling ability in these conditions is consistent with this interpretation.

One final alternative explanation with respect to choice can-

not be ruled out. In discussing the differences between the choice and no-choice conditions of this experiment, it must be pointed out that no-choice subjects received continuous feedback, whereas choice subjects had the option of choosing interval feedback. The differences in the two conditions may have been due to differences between continuous and interval feedback rather than between the options of choosing or not choosing feedback. Future research should contain a no-choice group provided with interval feedback to avoid this confound. Although such a group was not included in this experiment, we did notice that 36 of the subjects (40%) in the choice condition chose feedback on every trial, which is comparable to receiving the continuous feedback of the no-choice condition.

In conclusion, the results of this study suggest that when performance is evaluated on more than one dimension, it may be useful to provide feedback separately on all important dimensions and allow performers the freedom to choose feedback on each dimension to reduce the possibility of redundancy and minimize the amount of time needed to receive and evaluate feedback.

References

- Ammons, R. B. (1956). Effects of knowledge of performance: A survey and tentative theoretical formulation. *Journal of General Psychology*, 54, 279-299.
- Ashford, S. F., & Cummings, L. L. (1983). Feedback as an individual resource: Personal strategies for creating information. *Organizational Behavior and Human Performance*, 32, 370-398.
- Bandura, A., & Cervone, D. (1983). Self-evaluative and self-efficacy mechanisms governing the motivational effects of goal systems. *Journal of Personality and Social Psychology*, 45, 1017-1028.
- Becker, L. J. (1977). The joint effect of feedback and goal setting on performance: A field study of residential energy conservation. *Journal of Applied Psychology*, 63, 428-433.
- Campion, M. A., & Lord, R. G. (1982). A control systems conceptualization of the goal-setting and process. *Organizational Behavior and Human Behavior*, 30, 265-287.
- Earley, P. C. (1986). *An examination of the mechanisms underlying the relation of feedback to performance*. Unpublished manuscript, Claremont McKenna College, Department of Psychology, Claremont, CA.
- Furness, E. L., & Boyd, G. A. (1959). Three hundred thirty-five real spelling demons for college students. *College English*, 20, 294-295.
- Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, 64, 349-371.
- Komaki, J., Barwick, K. D., & Scott, L. R. (1978). A behavioral approach to occupational safety: Pinpointing and reinforcing safe performance in a food manufacturing plant. *Journal of Applied Psychology*, 64, 424-445.
- Larson, J. R., Jr. (1984). The performance feedback process: A preliminary model. *Organizational Behavior and Human Performance*, 33, 42-76.
- Locke, E. A., Cartledge, N., & Koeppe, J. (1968). Motivational effects of knowledge of results: A goal-setting phenomenon. *Psychological Bulletin*, 74, 474-485.
- Locke, E. A., Shaw, K. N., Saari, L. M., & Latham, G. P. (1981). Goal setting and task performance: 1969-1980. *Psychological Bulletin*, 90, 125-152.
- Matsui, T., Okado, A., & Kakuyama, T. (1982). Influence of need for achievement on goal setting, performance, and feedback effectiveness. *Journal of Applied Psychology*, 67, 645-648.
- Payne, R. B., & Hauty, G. T. (1955). The effect of psychological feedback on work decrement. *Journal of Experimental Psychology*, 50, 343-351.
- Rosenberg, M. (1965). *Society and the adolescent self image*. Princeton, NJ: Princeton University Press.
- Taylor, M. S., Fisher, C. D., & Ilgen, D. R. (1984). Individuals' reactions to performance feedback in organizations. In K. Rowland & G. R. Ferris (Eds.), *Research in Personnel and Human Resources Management* (Vol. 2, pp. 231-272). Greenwich, CT: JAI Press.
- Winer, B. J. (1962). *Statistical principles in experimental design*. New York: McGraw-Hill.

Received August 21, 1986

Revision received November 20, 1986

Accepted November 21, 1986 ■

Effects of Goals and Feedback on Performance in Groups

Tamao Matsui, Takashi Kakuyama, and Mary Lou Uy Onglatco
Rikkyo University, Tokyo, Japan

In the first study, 26 undergraduate pairs and 52 individuals worked on a perceptual speed task for 20 min to win prizes based on performance. The pairs set group goals and individual goals to be attained, whereas the individuals set only individual goals. Despite the equal levels of individual goals set, goal acceptance and performance were significantly higher for the pairs than for the individuals. A stepwise hierarchical regression analysis supported the contributions of goal acceptance and group goals to performance. In the second study, 50 undergraduate pairs were assigned a goal to be attained as teams on a perceptual speed task lasting 15 min. Group and individual task feedback, given after 7½ min of work, significantly improved performance only for those subjects who were below target for either group or individual feedback, yielding interaction effects on performance. The implications of the findings for group goal setting, social loafing, and organizational effectiveness are discussed.

In the last two decades, knowledge of the effects of individual goal setting on task motivation has accumulated (see Locke, Shaw, Saari, & Latham, 1981, for a review). However, the effects of group goal setting have rarely been studied.

Studies so far have dealt with two key attributes of group goals. One is goal clarity. Cohen (1959), Ishida (1980), Latham and Kinne (1974), and Watson (1983) found that specific goals led to better group performance than unspecified, vague goals. The other attribute is goal difficulty. Latham and Yukl (1975), Steers and Porter (1974), and Zander and Newcomb (1967) found that groups performed better if their goals were difficult than if they were easy. Yet, the question remains as to which type of goal, group or individual, would lead to better performance.

The present article reports two studies of the effects of goals and feedback on performance in groups. The first study predicted that subjects would perform better if the ultimate standard were group performance (referred to as group goal setting) than if the ultimate standard were individual performance (referred to as individual goal setting). If subjects set a goal to be attained as a team, and define their contribution to the attainment of the goal in terms of individual goals, they actually set two types of goals: a group goal and an individual goal. It is obvious that the group goal is much higher than any one individual goal under additive task demands (Steiner, 1972). Subjects naturally will see that the group goal is not their goal alone and is impossible to attain individually. Yet, Horwitz's (1954)

classic study demonstrated that group goals aroused motivational forces in group members that were similar to those aroused by individual goals. Erez and Zidon (1984) and Locke (1982) found that even impossible goals had positive effects on performance if the goals were accepted. Thus, subjects with group goals should strive, beyond the point of reaching their individual goal levels, to get as close as possible to their group goal, resulting in higher goal difficulty levels and thus higher performance levels than in the case in which subjects set individual goals alone.

In addition to increasing goal difficulty levels, group goal setting should enhance acceptance of individual goals. In group goal setting, subjects should develop a sense of shared responsibility for the attainment of their individual goals. This would motivate them to exert extra effort so that their performance would not cause the failure of the group. Pepitone (1952) found that members of groups felt responsible to their groups if they performed an important role in the groups. Erez, Earley, and Hulin (1985) found a positive relation between goal acceptance and performance. Thus, group goal setting should lead to higher performance than individual goal setting alone wherein goal acceptance is higher.

One possible qualification to these hypotheses may lie in the phenomenon of social loafing. Studies on social loafing (e.g., Albanese & Van Fleet, 1985; Harkins, Latané, & Williams, 1980; Harkins & Petty, 1982; Ingham, Levinger, Graves, & Peckham, 1974; Kerr, 1983; Kerr & Bruun, 1981, 1983; Latané, Williams, & Harkins, 1979; Williams, Harkins, & Latané, 1981) have presented various examples of motivation loss in groups. Latané et al. (1979) found that when subjects were asked to clap as loudly as possible in groups of differing sizes, the amount of noise per person was a negative function of group size. Note, however, that these subjects were not allowed to set any specific goals. They were merely told to do their best. Goal setting studies found that specific, challenging goals trigger maximum effort, but unspecified, vague goals like merely saying "do your best" do not (see Locke et al., 1981, for a review). People having no specific goals to attain as a group would free-

The authors would like to express their appreciation to Akinori Okada and Tadashi Imaizumi for their suggestions for statistics, and to the journal editor and two anonymous reviewers for their helpful comments on an earlier version.

Takashi Kakuyama is now at Tokyo International University in Shiki, Japan, and Mary Lou Uy Onglatco is now at Fujitsu Company, Limited, in Tokyo.

Correspondence concerning this article should be addressed to Tamao Matsui, who is now at Surugadai University, Hanno, Saitama 357, Japan.

ride on others' efforts (Kerr & Bruun, 1983) or lose motivation if they found their partners free-riding on them (Kerr, 1983).

Study 1

Method

Subjects

A total of 104 undergraduates (74 men and 30 women), enrolled in an industrial relations course, served as subjects. They were chosen from a pool of 211 students on the basis of pretest ability scores.

A numerical counting task was used to measure perceptual speed. Subjects were asked to count the frequency of a number that corresponded to a designated number to the left of each row of 50 numbers. Numbers within and across rows were completely randomized. Responses were recorded in a blank space to the right of each row of numbers.

A week before the experiment, the 211 students worked on the pretest, which lasted 2 min. They were told to do their best. After 2 min had elapsed, subjects were asked to count and record the number of correct responses, which served as the ability score.

The experimental trial was conducted with 52 group goal subjects (26 teams of 2 persons each) and 52 individual goal subjects, chosen from the pool of the 211 students (see ahead). The task was numerical counting similar to that used in the pretest. The trial lasted 20 min.

Experimental Conditions

The 211 students were divided in two groups with similar ability score distributions. One group ($n = 105$) served as a pool for group goal subjects, and the other group ($n = 106$), as a pool for individual goal subjects. The group goal and individual goal subjects were chosen from the pools so that the differences of team ability scores (i.e., the sum of ability scores for the two teammates) across the teams was smallest and the number of subjects and the distributions of ability scores were exactly equal for the two goal conditions. This was done to minimize the potential differences in the probability of winning prizes for the two goal conditions.

Group Goal Condition

The highest scoring student in the group goal subject pool was matched with the lowest scoring student, the second highest scoring student with the second lowest scoring student, and so forth. Although this procedure generated 52 two-member teams, it was found that only 26 teams of equal ability could be formed. Thus, these 26 teams (i.e., 52 subjects) were used for the group goal condition. The mean for ability scores was 10.8 ($SD = 2.29$). The ability scores for the two teammates totaled 21 or 22. The differences between the two teammates in ability scores within teams ranged from 1 to 8.

Subjects were informed that their team would win a prize if their team score (i.e., the sum of performance for the two teammates) fell within the highest six team scores. The prize had a value of about \$8. They were told that the teams were competing on the basis of equal team ability scores (i.e., 21 or 22). Thus, the probability of winning the prize was 23%. Team members were told to discuss and specify (a) the score the team would attempt to achieve (i.e., a group goal) and (b) each member's contribution to the attainment of the group goal (i.e., an individual goal).

After the group and individual goals were set, subjects were asked to move to assigned seats in a separate room where they worked on the experimental trial together with the individual goal subjects. The seats were completely randomized so that team members were not sitting together and could not encourage one another during the trial.

Individual Goal Condition

In all, 52 students were chosen from the individual goal subjects pool as individual goal subjects. The subjects were then classified into high- and low-ability categories. The 26 subjects whose ability scores matched those of the high-scoring group goal subjects were classified as the high-ability category, and the remaining subjects as the low-ability category. The ability scores ranged from 11 to 16 with a mean of 12.9 ($SD = 1.05$) for the high-ability category, and from 7 to 10 with a mean of 8.7 ($SD = .90$) for the low-ability category. The overall mean was 10.8, the same as that for the group goal subjects. To minimize the differences in the potential probability of winning a prize, subjects competed on the basis of their classification of either high- or low-ability categories. Subjects were informed of the category to which they belonged through their code in the questionnaire.

Subjects were told that they would win a prize if their score fell within the highest six individual scores in their respective categories. The prize had a value of about \$4. Thus, the probability of winning the prize was also 23% for both high- and low-ability subjects. Subjects were then asked to specify the number of rows they sought to attain to win the prize (i.e., their individual goal). After this, subjects moved to assigned seats in a separate room where they worked on the experimental task together with the group goal subjects.

Measures

Attractiveness of incentive. Subjects were asked to indicate the extent to which they found the incentive attractive. A 5-point scale, ranging from *unattractive* (1) to *extremely attractive* (5), was used for this purpose.

Desire to win. Subjects were asked to indicate their desire to win the prize on a 5-point scale. The scale ranged from *strongly undesirable* (1) to *strongly desirable* (5).

Goal acceptance. This measure was obtained by asking subjects to indicate the extent to which they would try to attain their individual goal. A 5-point scale, ranging from *will definitely not try* (1) to *will definitely try* (5), was used for this purpose.

Probability of success. Both group and individual goal subjects were asked to indicate their probability of winning the prize as a team, and personally, respectively. This was rated from 0% to 100%, with 10% intervals.

Performance. The number of correct answers made by subjects in the experimental trial was used as the index of performance.

Results

Although subjects worked as teams in the group goal condition, analyses were based on individuals' data. Table 1 shows means and standard deviations of main variables for the two goal conditions.

There are no significant differences between the two conditions in attractiveness of incentive, desire to win, and probability of success. These findings support the equality of motivational antecedents for the two goal conditions.

The performance mean is significantly higher for the group goal condition than for the individual goal condition, supporting the prediction. This finding suggests substantial effects of group goals on performance, inasmuch as there was no difference between the two goal conditions in ability level.

The individual goal means are not different for the two goal conditions. It is not surprising that the difference between the group goal and the individual goal for the group goal condition was significant ($M = 221.8$ and $M = 107.9$, respectively),

Table 1
Predictor and Criterion Means and Standard Deviations
for Two Goal Conditions

Variable	Goal condition		<i>t</i> (102)
	Group	Individual	
Attractiveness of incentive			
<i>M</i>	3.6	3.5	<i>ns</i>
<i>SD</i>	1.11	1.12	
Desire to win			
<i>M</i>	3.4	3.1	<i>ns</i>
<i>SD</i>	0.99	1.15	
Probability of success			
<i>M</i>	36.3	38.4	<i>ns</i>
<i>SD</i>	17.10	18.26	
Individual goal			
<i>M</i>	107.9	107.2	<i>ns</i>
<i>SD</i>	18.15	21.01	
Goal acceptance			
<i>M</i>	4.4	3.9	3.15**
<i>SD</i>	0.69	0.73	
Group goal			
<i>M</i>	221.8		
<i>SD</i>	22.47		
Performance			
<i>M</i>	114.9	105.5	2.41*
<i>SD</i>	20.73	18.48	

Note. Group *n* = 52; individual *n* = 52.

* *p* < .05. ** *p* < .01.

$t(50) = 42.86, p < .01$. Of more relevance, the group goals are significantly higher than the sum of the individual goals for the two teammates (for group goal and the sum of the individual goals, $M = 221.8$ and $M = 215.7$, respectively; $SD = 21.62$; for the difference, $M = 6.1$, $SD = 6.66$, $t[25] = 4.54, p < .01$, for correlated sample). These findings suggest that group goal subjects strove toward higher goals than did individual goal subjects.

Table 2 summarizes the means and standard deviations of individual goal and performance for high- and low-ability subjects for the two conditions. Despite the equal levels of individual goals for the two goal conditions as shown in Table 1, high and low group goal subjects exceeded their individual goals, although not significantly for low-ability subjects. For the total, performance was significantly higher than individual goal.

In contrast, high individual goal subjects did not reach their individual goal levels, whereas low individual goal subjects strove only up to their individual goal levels. For the total, performance did not differ from the individual goal. These findings suggest that goal acceptance was higher for group goal subjects than for individual goal subjects. In fact, the goal acceptance mean is significantly higher for the group goal condition than for the individual goal condition ($M = 4.4$ and $M = 3.9$, respectively), $t(102) = 3.15, p < .01$.

To determine the contributions of ability, individual goals, goal acceptance, goal conditions, and their interactions to performance, a stepwise hierarchical regression analysis was calculated across the two conditions. This analysis used performance

Table 2
Means and Standard Deviations of Individual Goals
and Performance for Two Ability Categories
for Two Goal Conditions

Ability category	Individual goal	Performance	<i>t</i> (25)
Group goal			
High (<i>n</i> = 26)			
<i>M</i>	118.0	128.6	2.70*
<i>SD</i>	17.73	17.14	
Low (<i>n</i> = 26)			
<i>M</i>	97.7	101.2	1.24
<i>SD</i>	11.71	13.81	
Total (<i>n</i> = 52)			
<i>M</i>	107.9	114.9	2.88 ^a **
<i>SD</i>	18.15	20.73	
Individual goal			
High (<i>n</i> = 26)			
<i>M</i>	121.4	117.4	1.13
<i>SD</i>	17.28	13.55	
Low (<i>n</i> = 26)			
<i>M</i>	93.0	93.6	0.19
<i>SD</i>	13.43	14.71	
Total (<i>n</i> = 52)			
<i>M</i>	107.2	105.5	0.70 ^a
<i>SD</i>	21.01	18.48	

^a *df* = 51.

* *p* < .05. ** *p* < .01.

as the dependent variable. Independent variables were, in order, ability, individual goal, goal acceptance, goal conditions (1 for the group goal condition, and 0 for the individual goal condition), and their interaction terms. Results are summarized in Table 3.

As shown in Table 3, ΔR^2 for Step 2 is significant. Thus, individual goal explains performance variance after removing ability. ΔR^2 is also significant for Steps 3 and 4. Goal acceptance explains performance variance after removing both ability and individual goal. Group goal explains performance variance even after all other effects were removed, suggesting the potent effects of group goals on performance.

Additional analyses were conducted to determine whether experimental conditions affected the results. The 26 teams had

Table 3
Regression Results for Performance

Step	Variable	R^2	ΔR^2	<i>F</i> for ΔR^2
1	Ability	.52	—	—
2	Individual goal	.55	.03	6.71*
3	Goal acceptance	.58	.03	7.14**
4	Goal condition	.61	.03	7.62**
5	Interaction (2 × 3)	.61	.00	0.99
6	Interaction (2 × 4)	.62	.01	2.55
7	Interaction (3 × 4)	.62	.00	1.80
8	Interaction (2 × 3 × 4)	.62	.00	0.22

* *p* < .05. ** *p* < .01.

differing combinations of male and female students. Thus, the effect of sex differences in the pairs on performance was tested. Group goal subjects were classified into male-male, female-female, or male-female pair category. Performance means were computed for the three categories. The performance means were 111.8 ($n = 22$, $SD = 21.75$), 106.8 ($n = 6$, $SD = 12.21$), and 119.8 ($n = 24$, $SD = 20.28$), respectively. There were no significant differences among the three performance means, $F(2, 49) = 1.35$. Because ability levels were equivalent as well ($M = 10.8$, $SD = 2.21$; $M = 10.7$, $SD = 1.48$; and $M = 10.8$, $SD = 2.56$, respectively), sex differences in the pairs did not affect performance.

To equalize the ability levels of groups, the highest ability subject was paired with the lowest ability subject, the second highest ability subject with the second lowest ability subject, and so forth. Thus, the ability levels of high-ability subjects were different for the teams. In such a situation, the group performance could have largely reflected the ability levels of high-ability subjects. To test this possibility, a stepwise, hierarchical regression was computed for group performance. Independent variables were, in order, the ability score of high-ability subjects, the ability score of low-ability subjects, and the interaction term. If the variance in group performance is unduly accounted for by the ability level of high-ability subjects, the entry of the score of low-ability subjects would not increase the value of R^2 . The results indicated significant increment in R^2 ($R^2 = .08$, $R^2 = .36$, and $R^2 = .37$, respectively; for the increment in R^2 , $F[1, 23] = 10.41$, $p < .01$, for the second step, and $F[1, 22] = .12$, ns , for the third step), thereby precluding the possibility.

Low-ability subjects worked alone in the individual goal condition, whereas they worked with high-ability subjects in the group goal condition. This could have led to increased performance for low-ability subjects in the group goal condition (for low-ability subjects' performance, $M = 101.2$, $SD = 13.81$, for the group goal condition, and $M = 93.8$, $SD = 14.71$, for the individual goal condition; $t[50] = 1.88$, $p < .10$). Laughlin and Johnson (1966) found that lower ability subjects performed better when they worked with higher ability subjects than when they worked alone or with those of the same ability levels. To test this possibility, a stepwise hierarchical regression analysis was computed for low-ability subjects' performance. Independent variables were, in order, the ability score of low-ability subjects, the ability score of high-ability subjects, and their interaction term. Results indicated that neither the ability score of high-ability subjects nor the interaction term increased R^2 ($R^2 = .18$, $R^2 = .26$, and $R^2 = .27$, respectively; for the increment in R^2 , $F[1, 23] = 2.29$, ns , for the second step, and $F[1, 22] = .49$, ns , for the third step), thereby suggesting the minimal influence of high-ability subjects on low-ability subjects' performance. Laughlin and Johnson (1966) used a problem solving task in which solutions were reached on the basis of group discussion. On the other hand, the present subjects worked independently. Thus, high-ability subjects did not affect the performance of low-ability subjects.

In his review article, Eden (1984) noted that the high expectation of supervisors has a positive influence on the performance of their subordinates. When subjects set goals, they were told whether they belonged in the high- or low-ability category. Thus, low-ability subjects might expect a larger contribution to

the group performance from high-ability subjects. It is likely that the lower the ability levels of subjects, the higher their expectation that high-ability subjects would perform better. This could have positively affected the performance of high-ability subjects (for high-ability subjects' performance, $M = 128.6$, $SD = 17.14$, for the group goal condition, and $M = 117.4$, $SD = 13.55$, for the individual goal condition; $t[50] = 2.56$, $p < .05$). To test this possibility, a stepwise hierarchical regression was computed for the performance of high-ability subjects in the group goal condition. The independent variables were, in order, the ability score of high-ability subjects, the ability score of low-ability subjects, and their interaction term. This analysis did not yield significant increment in R^2 for either the score of low-ability subjects or the interaction term ($R^2 = .27$, $R^2 = .35$, and $R^2 = .38$, respectively; for the increment in R^2 , $F[1, 23] = 2.56$, ns , for the second step, and $F[1, 22] = 1.04$, ns , for the third step). These findings indicate that pairing high- and low-ability subjects, and informing subjects about their ability category did not bias the results.

Discussion

This study found that group goal setting led to higher performance than did individual goal setting. In addition, it suggests that group goal setting would facilitate performance through two mechanisms. One is increased goal difficulty. The individual goal levels were equal for the two conditions. Group goals were higher than individual goals for either teammate, and even higher than their sum. These findings suggest that group goal subjects strove, in effect, for higher goals than did individual goal subjects, resulting in higher performance.

The other mechanism is enhanced acceptance of individual goals. Despite the equal levels of individual goals for the two conditions, group goal subjects exceeded their individual goals, whereas individual goal subjects strove only up to their individual goal levels. This suggests that group goal subjects accepted their goals to a greater degree than did individual goal subjects, and in fact, the goal acceptance mean was significantly higher for the group goal condition than for the individual goal condition.

The study so far demonstrates the effects of group goals and individual goals on performance. However, goal-setting studies with individuals showed that feedback is a necessary accompaniment of goals to improve subsequent performance (Erez, 1977; Locke et al., 1981; Strang, Lawrence, & Fowler, 1978). This suggests the need for investigating the relation of goals and feedback in improving performance in group goal setting. Thus, the second study was conducted to investigate this relation.

Subjects can receive two types of task feedback in group goal setting. One is feedback on their own performance, and the other, feedback on group performance. Thus, one important issue concerning feedback in group goal setting is the mechanisms by which the two types of task feedback affect performance. It was found in the 1960s that feedback containing both individual and group performance information was more effective than that containing group performance information alone (e.g., Zajonc, 1962; Zander & Wolfe, 1964). However, these studies did not consider the concept of goal setting. In

addition, the mechanisms by which feedback containing both types of performance information improved performance was not identified.

Peak (1955) identified that differences in the way one feels about a present state and an imagined state are the most important determinants of motive structure, and account for the persistence and intensity of motivated behavior. More recently, Campion and Lord (1982) conceptualized the relations among goals, feedback, and performance on the basis of the control systems model that was initially developed by Powers (1973). According to Campion and Lord, when subjects receive feedback, a referent signal (goal) and sensor signal (task feedback) are compared by a "comparator." If a sufficiently large discrepancy exists, some form of remedial action (e.g., goal revision as cognitive change or increased effort as behavioral change) is triggered.

Many studies have supported the control systems model in individual goal setting. Campion and Lord (1982) found that students who failed in attaining test goals increased effort as a function of the magnitudes and frequencies of failure. Matsui, Okada, and Inoshita (1983) found that when subjects were assigned goals and received feedback midway through their work, those subjects below target improved performance, whereas those who were on target merely maintained their previous levels of performance. Bandura and Cervone (1983) found that feedback was most effective for subjects who were below target, were dissatisfied with their past performance, and had a high level of perceived self-efficacy. These findings suggest that a negative goal discrepancy in the comparison process is essential for task feedback to improve performance.

If subjects set group goals and individual goals and receive group and individual task feedback during work, the comparison process would take place for the two types of feedback. If subjects found negative goal discrepancy in either comparison process, they would try to minimize it. In such a situation, subjects can be categorized into four groups, as shown in Figure 1. The *G* and the *I* refer to group and individual task feedback. The feedback is positive when the group or individual performance is either on or above the specified target. It is negative when the same performance is below the target. *P* indicates the level of performance.

Subjects in Category A are on target for both group and individual feedback. They receive positive task feedback from the two comparison processes. They can be expected merely to maintain their previous levels of efforts, resulting in no change in performance. Subjects in Category B, on the other hand, are on target individually, but their group is below target. They receive positive feedback from the comparison process for the individual task feedback and negative feedback from the comparison process for the group task feedback. They are expected to improve their performance to minimize the group goal discrepancy. Subjects in Category C have group performance on target, but they are below target individually. They receive positive feedback from the comparison process for the group task feedback, and negative feedback from the individual task feedback. They are expected to improve performance to minimize the individual goal discrepancy. Finally, subjects in Category D are below target for both group and individual task feedback. They receive negative feedback from the two comparison processes.

		GROUP FEEDBACK	
		On target	Below target
INDIVIDUAL FEEDBACK	On target	(Category A) <i>G</i> : positive <i>I</i> : positive <i>P</i> : no change	(Category B) <i>G</i> : negative <i>I</i> : positive <i>P</i> : improve
	Below target	(Category C) <i>G</i> : positive <i>I</i> : negative <i>P</i> : improve	(Category D) <i>G</i> : negative <i>I</i> : negative <i>P</i> : improve

Figure 1. Subject categories, nature of feedback, and predicted performance changes. (*G* = group task feedback; *I* = individual task feedback; and *P* = predicted performance changes after feedback).

They are expected to improve performance to minimize either goal discrepancy. In short, the task feedback would improve performance only for those below target for one or more sources of task feedback. Thus, the second study hypothesized the interaction effects of group and individual task feedback on performance. In addition, it specified that the interaction must be complementary, not additive in nature.

It may also be argued that subjects in Category B can lose motivation by finding their partners free-riding on them. Similarly, subjects in Category C can lose motivation by free-riding on their partners. Kerr (1983) found motivation loss due to the "sucker effect," when group members reduced efforts if they had capable partners who free-rode on their efforts. Kerr and Bruun (1983) found motivation loss due to the free-rider effect when group members exerted less effort as the perceived dispensability of their efforts for group success increased. Note, however, that these subjects worked under disjunctive or conjunctive task demands in which the standard was the performance of more capable or less capable subjects. Thus, the efforts of more capable or less capable subjects were dispensable, with no reason to work hard. It is interesting to note that Kerr and Bruun (1983) failed to find the free-rider effect when subjects worked under additive task demands. In addition, subjects were not allowed to set either group or individual goals. If the task demand is an additive one, and subjects set both group and individual goals that are attainable only if they worked hard, motivation loss due to the sucker or free-rider effect would be prevented.

Study 2

Method

Subjects, Task, and Procedure

Subjects were 89 male and 11 female university students enrolled in an industrial relations course. A perceptual speed task was used that was similar to that in the first study.

Before the experimental trial, a 2-min pretest using the same task as

Table 4
Means, Standard Deviations, and Intercorrelations of Major Variables

Variable	M	SD	1	2	3	4	5	6
1. Ability	11.3	2.88	—					
2. Probability of success	67.5	22.22	.09	—				
3. Goal acceptance	4.0	0.94	.15	.18*	—			
4. Individual goal	80.4	12.94	.89**	.12	.19*	—		
5. Prefeedback performance	40.8	8.48	.77**	.10	.13	.72**	—	
6. Postfeedback performance	44.0	8.48	.72**	.11	.16	.70**	.87**	—

Note. N = 100.
* $p < .05$, one-tailed. ** $p < .01$.

that used in the experimental trial was conducted. Subjects were instructed to do their best. After 2 min had elapsed, subjects were asked to count and record the number of correct answers, which then served as an ability score.

After the pretest, the highest scoring subject was matched with the lowest scoring subject, the second highest scoring subject with the second lowest scoring subject, and so forth, resulting in 50 teams. The sum of the ability scores for the two teammates ranged from 21 to 24. The difference between the two teammates in ability score in teams ranged from 1 to 8.

Subjects were asked to sit together with their team-mates, and were informed what their team ability score was. Subjects were told that there would be a trial lasting 15 min, with a team goal to be attained. The goal was to complete 160 rows as a team for the period. Subjects were then asked to discuss with their teammate the number of rows that each member would attempt in order to attain the assigned team goal.

After 7½ min of work, subjects were asked to stop, and were informed that one half of the total time had elapsed. Subjects then received feedback on both group and individual performance. After receiving feedback, subjects were asked to continue the task.

Measures

Probability of success. Subjects were asked to indicate the probability of attaining the goal as a team. The scale ranged from 0% to 100%, with 10% intervals.

Goal acceptance. This measure was obtained by asking subjects to indicate the extent to which they would try to attain the team goal. A 5-point scale, ranging from *will definitely not try* (1) to *will definitely try* (5), was used for this purpose.

Individual and group performance. The number of correct answers made by each subject in the experimental trial served as the index of individual performance. The sum of individual performance for the two teammates was the index of group performance.

Results

Although subjects worked as teams, the following analyses were conducted on the basis of individuals' data. Table 4 shows means, standard deviations, and intercorrelations of main variables.

The table indicates that the assigned goal was perceived by subjects as moderately difficult ($M = 67.5$, for probability of success), and was accepted by most of the subjects ($M = 4.0$, for goal acceptance). Postfeedback performance is significantly higher than prefeedback performance, which suggests that task feedback enhanced performance (for the difference between

postfeedback and prefeedback performance, $M = 3.2$, $SD = 4.34$; $t[99] = 7.35$, $p < 0.1$).

Table 5 shows the means of performance changes before and after feedback for the four subject categories shown in Figure 1. Subjects are classified as on target for group feedback if their prefeedback group performance was equal to or higher than 80 (i.e., one half of the assigned goal). Similarly, subjects are classified as on target for individual feedback if their prefeedback individual performance was equal to or higher than one half of the personal goals. The remaining subjects are classified as below target for group and individual feedback, respectively.

As shown in Table 5, no significant performance change occurs for Category A, $t(32) = 1.29$, ns , whereas postfeedback performance is significantly higher than prefeedback performance for the remaining three categories, $t(13) = 3.76$, $p < .01$; $t(24) = 5.22$, $p < .01$; and $t(27) = 6.22$, $p < .01$, respectively. The mean for Category A is significantly lower than that for the remaining three categories, $t(45) = 3.10$, $p < .01$, for Category B; $t(56) = 3.28$, $p < .01$, for Category C; and $t(59) = 2.87$, $p < .01$, for Category D. There are no significant differences among the means for Categories B, C, and D, $F(2, 64) = .96$, ns . These

Table 5
Means and Standard Deviations of Performance Changes for Four Subject Categories

Individual feedback	Group feedback	
	On target	Below target
On target	Category A $n = 33$ $M = 0.9$ $SD = 3.96$	Category B $n = 14$ $M = 5.4$ $SD = 5.13$
Below target	Category C $n = 25$ $M = 4.6$ $SD = 4.28$	Category D $n = 28$ $M = 3.6$ $SD = 2.98$

Note. On target for group (individual) feedback = subjects whose prefeedback group (individual) performance was equal to or higher than one half of group (individual) goal; Below target = remaining subjects. For the mean difference, $t(32) = 1.29$, ns , for Category A; $t(13) = 3.76$, $p < .01$, for Category B; $t(24) = 5.22$, $p < .01$, for Category C; and $t(27) = 6.22$, $p < .01$, for Category D.

findings are consistent with the predictions made in Figure 1, which suggest the interaction effects between the two types of task feedback on performance. In addition, the finding that the mean for Category D did not significantly differ from that for Categories B and C suggests that the effects of both types of task feedback were complementary, not additive, in nature.

Additional analyses were conducted to determine whether the free-rider effect (Kerr & Bruun, 1983) and the sucker effect (Kerr, 1983) had occurred. The free-rider effect is defined as reduced efforts due to dispensability of efforts for group success (Kerr & Bruun, 1983). This effect was thought to be possible for those subjects whose performance was below target, but group performance and partners' performance were on target. Thus, performance changes were computed for 23 subjects belonging to Category C whose partners were on target. If they intended to free-ride on their partners, postfeedback performance should be lower than prefeedback performance. However, postfeedback performance was significantly higher than prefeedback performance (for postfeedback and prefeedback performance, $M = 42.3$, $SD = 6.52$, and $M = 37.5$, and $SD = 6.04$, respectively; for the difference, $M = 4.8$, $SD = 4.36$, $t[22] = 5.19$, $p < .01$). Thus, there was no free-rider effect.

The sucker effect, on the other hand, is defined as reduced efforts of group members who have capable partners free-riding on their efforts (Kerr, 1983). This effect was thought to be possible for lower ability subjects whose performance was on target but whose partners' performance was below target. Thus, performance changes were computed for 10 lower ability subjects belonging to Category A, and 7 lower ability subjects belonging to Category B, whose partners' performance was below target. For the lower ability subjects belonging to Category A, postfeedback performance did not differ from prefeedback performance ($M = 42.8$, $SD = 4.26$, and $M = 43.5$, $SD = 4.67$, respectively; for the difference, $M = -0.7$, $SD = 2.63$, $t[9] = .84$, ns). For the lower ability subjects belonging to Category B, on the other hand, postfeedback performance was significantly higher than prefeedback performance ($M = 39.7$, $SD = 4.74$, and $M = 35.9$, $SD = 4.22$, respectively; for the difference, $M = 3.9$, $SD = 3.02$, $t[6] = 3.37$, $p < .05$). These findings indicate that the lower ability subjects did not lose motivation even when they learned that their higher ability partners were performing poorly. It is interesting that lower ability subjects belonging to Category A and those belonging to Category B reacted differently to the same information that their capable partners were below target. The former merely maintained their previous levels of performance because their group was on target, whereas the latter improved performance because their group was below target, thereby demonstrating the potent effects of group goals on task motivation.

To confirm the interaction effects of the two types of task feedback suggested in Table 4, a stepwise hierarchical regression analysis was calculated. The analysis used postfeedback performance as the dependent variable. Independent variables were, in order, prefeedback performance, individual goal discrepancy, group goal discrepancy, and the interaction of the two goal discrepancies. A score of 1 was assigned to those subjects previously classified as on target, and 0, to those classified as below target for individual goal discrepancy. A similar procedure was used for group goal discrepancy. Difference scores

Table 6

Regression Results for Postfeedback Performance

Step	Variable	R^2	ΔR^2	F for ΔR^2
1	Prefeedback performance	.76	—	—
2	Individual goal discrepancy	.76	.00	0.76
3	Group goal discrepancy	.76	.00	1.28
4	Interaction (2×3)	.78	.02	9.32*

* $p < .01$.

were not used as goal discrepancies because the correlation between the individual and group goal discrepancies was .50. With such a high correlation for the components of interaction terms, the variance accounted for by the interaction terms would be unduly small. The high correlation resulted from the facts that the individual and group goal discrepancies have in common individual prefeedback performance as a component, and that the teams consisted of only 2 subjects. In addition, it was thought that subjects' perceptions of below target, as well as the magnitudes of goal discrepancies, would trigger increased efforts. Table 6 summarizes results. It indicates that ΔR^2 is significant only for Step 4, substantiating the interaction effects of the two types of task feedback hypothesized.

Discussion and Concluding Remarks

This study supports the control systems model of task feedback (Campion & Lord, 1982) in group goal setting. If subjects receive both group and individual task feedback, the comparison process between goal and performance would take place for the two types of task feedback. If subjects find a negative goal discrepancy in either process, they would try to minimize it, resulting in improved performance. Subjects would maintain their previous levels of performance only if they found no discrepancies in the two comparison processes. Thus, the effects of group and individual task feedback on performance would be complementary, not additive in nature.

The findings from the study suggest that the effectiveness of task feedback in group goal setting will be maximized if the feedback involves both individual and group performance information. If feedback does not involve individual performance information, those subjects below target would not improve performance when their group performance was on target (equivalent to those in Category C in Figure 1). On the other hand, if the feedback does not involve group performance information, those subjects with their group performance below target would not improve performance when they are on target individually (equivalent to those in Category B in Figure 1). In contrast, if the feedback involves both group and individual task feedback, information of lower group (individual) progress would trigger increased effort to improve performance, despite information of higher individual (group) progress, resulting in maximum utilization of subjects' resources. Locke and Latham (1984) also argued that to ensure maximum performance, the performance of individuals and groups in relation to goals should be measured.

The present findings also have implications for the prevention of social loafing. It was found that subjects who were below target did not stop striving even after they learned that their group was on target. Feedback enabled them to identify their own low progress that resulted in dissatisfaction. Thus, they worked harder to catch up, which prevented them from free-riding on their partners' effort. Subjects who were on target did not stop striving even after they learned that their more capable partners were performing poorly, and did not show the sucker effect. This was possible because they had the group goal to be attained, and felt that their efforts were indispensable. Social loafing studies found that motivation loss was minimized when the tasks increased identifiability (Williams, Harkins, & Latané, 1981) or decreased dispensability (Kerr & Bruun, 1983) of efforts. Specific group goals and individual goals minimize dispensability of efforts. Feedback on both group and individual performance maximize identifiability of efforts. Thus, setting both specific group and individual goals with the provision of feedback on performance in relation to both goals would be an effective strategy to prevent motivation loss when people work together.

The implications of the present findings for organizations deserve mention. Organizations consist of many groups: departments, sections, task forces, and so forth. The findings suggest that having members work as teams with a specific team goal rather than as individuals with only individual goal increases productivity. In addition, to maximize the productivity of such teams, information and control systems (Lawler & Rhode, 1976) should contain information on team progress and individual progress. Although these suggestions are based on findings from the laboratory setting, which is much simpler than organizational settings, Latham and Lee (1986) found that goal-setting findings of laboratory studies were consistent with those of field studies.

Some methodological issues should be discussed. The possible effects of cultural attributes on goal acceptance might have limited the generalizability of the present findings. Although empirical evidence is scarce, it has been argued that the Japanese culture, including the present subjects, places its emphasis on fulfilling team expectations to a greater degree than do cultures of other countries like the United States. Assuming that this argument is reasonable, the cultural attributes emphasizing group attainment might have unduly enhanced goal acceptance and prevented the occurrence of free-rider or sucker effect. Locke, Latham, and Erez (in press) argued that the effect of group versus individual goal setting on goal acceptance may be moderated by cultural attributes.

Finally, the present studies relied on the simplest form of groups (i.e., two-member teams). This might have led to high goal acceptance. Such acceptance, however, could be reduced if group size is increased. Thus, replications of the present findings in different cultural contexts with larger groups would be worthwhile in future studies.

References

- Albanese, R., & Van Fleet, D. D. (1985). Rational behavior in groups: The free-riding tendency. *Academy of Management Review*, 10, 244-255.
- Bandura, A., & Cervone, D. (1983). Self-evaluative and self-efficacy mechanisms governing the motivational effects of goal systems. *Journal of Personality and Social Psychology*, 45, 1017-1028.
- Campion, M. A., & Lord, R. G. (1982). A control systems conceptualization of the goal-setting and changing process. *Organizational Behavior and Human Performance*, 30, 265-287.
- Cohen, A. R. (1959). Situational structure, self-esteem, and threat-oriented reactions to power. In D. Cartwright (Ed.), *Studies in social power*. Ann Arbor, MI: Institute for Social Research.
- Eden, D. (1984). Self-fulfilling prophecy as a management tool: Harnessing pygmalion. *Academy of Management Review*, 9, 64-73.
- Erez, M. (1977). Feedback: A necessary condition for the goal-performance relationship. *Journal of Applied Psychology*, 62, 624-627.
- Erez, M., Earley, P. C., & Hulin, C. L. (1985). The impact of participation on goal acceptance and performance: A two-step model. *Academy of Management Journal*, 28, 50-66.
- Erez, M., & Zidon, I. (1984). Effect of goal acceptance on the relationship of goal difficulty to performance. *Journal of Applied Psychology*, 69, 69-78.
- Harkins, S. T., Latané, B., & Williams, K. (1980). Social loafing: Allocating effort or taking it easy? *Journal of Experimental Social Psychology*, 16, 457-465.
- Harkins, S. T., & Petty, R. E. (1982). Effects of task difficulty and task uniqueness on social loafing. *Journal of Personality and Social Psychology*, 43, 1214-1229.
- Horwitz, M. (1954). The recall of interrupted group tasks: An experimental study of individual motivation in relation to group goals. *Human Relations*, 7, 3-38.
- Ingham, A. G., Levinger, G., Graves, J., & Peckham, V. (1974). The Ringelmann effect: Studies of group size and group performance. *Journal of Experimental Social Psychology*, 10, 371-384.
- Ishida, H. (1980). The effects of varied clarity of group goal and substeps upon group problem solving. *Japanese Journal of Experimental Social Psychology*, 19, 119-125.
- Kerr, N. L. (1983). Motivation losses in small groups: A social dilemma analysis. *Journal of Personality and Social Psychology*, 45, 819-828.
- Kerr, N. L., & Bruun, S. E. (1981). Ringelmann revised: Alternative explanations for the social loafing effect. *Personality and Social Psychology Bulletin*, 7, 224-231.
- Kerr, N. L., & Bruun, S. E. (1983). Dispensability of member effort and group motivation losses: Free-rider effects. *Journal of Personality and Social Psychology*, 44, 78-94.
- Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, 37, 822-832.
- Latham, G. P., & Kinne, S. B. (1974). Improving job performance through training in goal setting. *Journal of Applied Psychology*, 59, 187-191.
- Latham, G. P., & Lee, T. W. (1986). Goal setting. In E. A. Locke (Ed.), *Generalizing from laboratory to field settings* (pp. 101-117). Lexington, MA: Heath.
- Latham, G. P., & Yukl, G. A. (1975). A review of research on the application of goal setting in organizations. *Academy of Management Journal*, 18, 824-845.
- Laughlin, P. R., & Johnson, H. H. (1966). Group and individual performance on a complementary task as a function of initial ability level. *Journal of Experimental Social Psychology*, 2, 407-414.
- Lawler, E. E., & Rhode, J. G. (1976). *Information and control in organizations*. Santa Monica, CA: Goodyear.
- Locke, E. A. (1982). Relation of goal level to performance with a short work period and multiple goal levels. *Journal of Applied Psychology*, 67, 512-514.
- Locke, E. A., & Latham, G. P. (1984). *Goal setting for individuals*,

- groups, and organizations (Module). Chicago: Science Research Associates.
- Locke, E. A., Latham, G. P., & Erez, M. (in press). The determinants of goal commitment. *Academy of Management Review*.
- Locke, E. A., Shaw, K. N., Saari, L. M., & Latham, G. P. (1981). Goal-setting and task performance: 1968-1980. *Psychological Bulletin*, 90, 125-152.
- Matsui, T., Okada, A., & Inoshita, O. (1983). Mechanism of feedback affecting task performance. *Organizational Behavior and Human Performance*, 31, 114-122.
- Peak, H. (1955). Attitude and motivation. *Nebraska Symposium on Motivation* (Vol. 3, pp. 149-188). Lincoln: University of Nebraska Press.
- Pepitone, E. (1952). *Responsibility to group and its effects on the performance of members*. Unpublished doctoral dissertation, University of Michigan.
- Powers, W. T. (1973). Feedback: Beyond behaviorism. *Science*, 179, 351-356.
- Steers, R. M., & Porter, L. W. (1974). The role of task-goal attributes in employee performance. *Psychological Bulletin*, 81, 434-452.
- Steiner, I. D. (1972). *Group process and productivity*. New York: Academic Press.
- Strang, H. R., Lawrence, E. C., & Fowler, P. C. (1978). Effects of assigned goal level and knowledge of results on arithmetic computation: A laboratory study. *Journal of Applied Psychology*, 63, 446-450.
- Watson, C. (1983, August). *Motivational effects of feedback and goal setting on group performance*. Paper presented at the meeting of the American Psychological Association, Anaheim, CA.
- Williams, K., Harkins, S., & Latané, B. (1981). Identifiability as a deterrent to social loafing: Two cheering experiments. *Journal of Personality and Social Psychology*, 40, 303-311.
- Zajonc, R. B. (1962). The effects of feedback and probability of group success on individual and group performance. *Human Relations*, 15, 149-161.
- Zander, A., & Newcomb, T., Jr. (1967). Group levels of aspiration in United Fund campaigns. *Journal of Personality and Social Psychology*, 6, 157-162.
- Zander, A., & Wolfe, D. (1964). Administrative rewards and coordination among committee members. *Administrative Science Quarterly*, 9, 50-69.

Received February 3, 1986

Revision received December 1, 1986

Accepted January 26, 1987 ■

Task Complexity as a Moderator of Goal Effects: A Meta-Analysis

Robert E. Wood

Australian Graduate School of Management
University of New South Wales
Kensington, New South Wales, Australia

Anthony J. Mento

Loyola College in Maryland

Edwin A. Locke

University of Maryland

Much evidence exists that supports the use of goal setting as a motivational technique for enhancing task performance; however, little attention has been given to the role of task characteristics as potential moderating conditions of goal effects. Meta-analysis procedures were used to assess the moderator effects of task complexity for goal-setting studies conducted from 1966 to 1985 ($n = 125$). The reliability of the task complexity ratings was .92. Three sets of analyses were conducted: for goal-difficulty results (hard vs. easy), for goal specificity-difficulty (specific difficult goals vs. do-best or no goal), and for all studies collapsed across goal difficulty and goal specificity-difficulty. It was generally found that goal-setting effects were strongest for easy tasks (reaction time, brainstorming), $d = .76$, and weakest for more complex tasks (business game simulations, scientific and engineering work, faculty research productivity), $d = .42$. Implications for future research on goal setting and the validity of generalizing results are discussed.

Twenty years of empirical research has established that specific, challenging goals lead to higher levels of task performance than no goals, vague goals, or easy goals (Locke, Shaw, Saari, & Latham, 1981; Pinder, 1984). Interest is now turning to identifying the theoretical limits of goal setting, for example, the variables that moderate the positive performance effects of goals (e.g., Austin & Bobko, 1984; Locke et al., 1981; Naylor & Ilgen, 1984).

One variable of potential importance to the theory of goal setting is task complexity. Tasks are an integral part of all studies of human performance, and task characteristics have been suggested as moderators in a diverse range of areas, including job design (Hackman & Oldham, 1980), personnel selection (e.g., Peterson & Bownas, 1982), information processing and decision making (e.g., Streufert & Streufert, 1978), and psychomotor activities (Fleishman, 1975). Locke et al. (1981) have speculated that goals will have less direct effects due to effort, attention, and persistence on complex tasks, and that indirect effects due to strategy development will become more important to performance. Further, Wood (1985) has argued that specific, challenging goals may lead to suboptimal search procedures on complex tasks (e.g., Baumler, 1971; Huber, 1985). From this literature we hypothesize that goals, on the average, will have less pronounced performance effects on complex tasks than on simple tasks.

The study of characteristics such as complexity has suffered

from a lack of standardization in definition and a confounding of individual and task characteristics (Fleishman, 1975; Hackman, 1969; Weick, 1965; Wood, 1986). As a result, there has been a lack of integration of the evidence for task effects from studies in different areas and, in some cases, inconsistencies in results in a given area (Wood, 1986).

The problems of inconsistent results are evident in goal-setting studies that have considered the effects of task characteristics. Frost and Mahoney (1976) found that challenging goals led to higher performance than did easy goals on a simple task but not on a complex task. Jackson and Zedeck (1982), however, reported no such differences for two tasks of differing complexity. Frost and Mahoney also found that subjects with specific, challenging goals outperformed subjects with nonspecific goals by more on the complex task than on the simple task. Baumler (1971), using a much more complex set of tasks, obtained exactly the opposite results.

The inconsistencies in results between these studies are probably due, in part, to different manipulations of task complexity. These were path-goal multiplicity (Frost & Mahoney, 1976), manual versus cognitive tasks (Jackson & Zedeck, 1982), and interdependence of decisions (Baumler, 1971). Moreover, in two of these three studies, the tasks used in the complex condition were not very complex. In the Frost and Mahoney study, the complex task was a jigsaw puzzle with all pieces painted the same color. In Jackson and Zedeck's study, subjects performing the complex task had to calculate floor covering requirements from a plan for a single-level, three-room building. Neither of these tasks approach the complexity of tasks such as professional counseling, investment decision making, or air-traffic control (Wood, 1986). Therefore, the moderating effects of tasks may not have been adequately tested in either of these studies because of weak manipulations of complexity.

Baumler (1971) used a more powerful manipulation of task

Funding for this research was provided to the second author through a Summer Research Grant from the Sellinger School of Business and Management at Loyola College.

Correspondence concerning this article should be sent via air mail to Robert E. Wood, Australian Graduate School of Management, University of New South Wales, P.O. Box 1, Kensington, New South Wales 2033, Australia.

complexity; however, his study only examined one goal attribute (specificity). Also, because the levels of goals assigned to subjects were not reported, the possibility that the lower performance for specific goals in the complex condition was due to subjects setting lower goals, cannot be ruled out. Therefore, to this point, no adequate test of task complexity as a moderator of goal effects has been conducted.

Of course, goal setting has been found to have positive performance effects on tasks of varying complexity, ranging from simple brainstorming tasks (e.g., Locke, 1967) to college course work (Locke & Bryan, 1968) to complex scientific work (Latham, Mitchell, & Dossett, 1978). However, this cannot be accepted as evidence that tasks do *not* moderate the goal performance effects of goals without considering the size of effects for tasks of differing complexity. It is possible, for the reasons mentioned earlier, that goals have a relatively large performance effect on simple tasks and a smaller, but significant, effect on complex tasks. If this were the case, then the correct test for the moderator effects of task complexity would be a comparison of the magnitude of effects between goal setting studies that have used simple tasks and those that have used complex tasks. Meta-analysis is a technique that allows such a comparison.

Within the last decade, meta-analysis techniques have been developed that permit the quantitative aggregation of results across studies (Hunter, Schmidt, & Jackson, 1982; Rosenthal, 1978). An advantage of meta-analysis for our current purposes is that the influence of moderator variables such as task complexity can be examined. This first requires that the results of all studies be converted to a common statistic, either d or r_{pb} (Hunter et al., 1982), so that results can be cumulated across studies.

After correcting for as many of the different sources of error variance as possible (i.e., sampling error, range restriction, reliability of measures, etc.), any remaining variance in the results can be tested to see if differences in size of effects between studies are due to the hypothesized moderator. If the majority of the variance in the results of studies is removed by the corrections for errors, then the argument for any moderator effect can be rejected without any further testing. Schmidt, Hunter, and Pearlman (1982) have suggested that if approximately 75% of the total variance across studies can be accounted for by sampling, then any apparent moderator effect is most likely due to capitalization on chance.

A test of the moderator hypothesis is the strength of the relationship between the study statistic and the moderator variable. This can be established by regressing the moderator variable on the study statistic (Mabe & West, 1982; Steel & Ovalle, 1984). With the regression approach, testing the significance of the beta weight is a test of the moderator effect. Because it allows for tests to be based on the total sample, this approach is less susceptible to chance effects than the subgroup approach described by Hunter et al. (1982). However, it does require at least an ordinal measure on the moderator for the purposes of regression analyses. Because task complexity is ordinal in nature, as distinct from other commonly used moderators—such as sex—which are nominal categories, this approach was considered appropriate for the moderator test in the current study. The validity of the results also depends on the reliability of the measure

of the moderator variable. The reliability of the task-complexity scores in the present study was very high ($r = .92$).

Thus, the purpose of the present study was to examine the moderating effects of task complexity across existing goal-setting studies, using a meta-analysis approach. Based on earlier speculations (Locke et al., 1981; Wood, 1985), larger effects should be found for studies using simple tasks and smaller effects for studies with more complex tasks. The specific hypotheses to be tested are as follows:

Hypothesis 1. The positive performance effects of specific and difficult goals (vs. do-best goals) will be greater on simple tasks than on complex tasks.

Hypothesis 2. The positive performance effects of difficult goals (vs. moderate or easy goals) will be greater on simple tasks than on complex tasks.

Method

To identify studies appropriate for the meta-analyses, we manually searched the *Psychological Abstracts* and the *Social Science Citation Index* and systematically reviewed the *Journal of Applied Psychology*, *Academy of Management Journal*, *Organizational Behavior and Human Performance*, and *Personnel Psychology* from January 1966 to December 1984 (see Appendix). Studies were excluded if an effect size could not be calculated.

In the goal-difficulty analysis, three studies were excluded because they contained an experimental artifact in the easy-goal condition that involved instructing subjects to stop working when the easy goal was reached. This instruction may serve to artifactually inflate the goal-difficulty–performance relationship. A number of studies were excluded from the goal difficulty and goal specificity–difficulty meta-analyses that used a within-subjects, as opposed to a between-subjects, experimental design. In discussing quantitative approaches to literature reviews, Green and Hall (1984) cautioned that it is incorrect and inappropriate to include data from a within-subjects design into a meta-analysis because effect sizes cannot be accurately computed. Unpublished studies known to the authors were included. A list of specific studies excluded from the meta-analyses can be found in Mento, Steel, and Karren (1987).

The remaining studies available for the meta-analyses included 72 studies of goal-difficulty effects (difficult vs. moderate or easy) and 53 studies of goal specificity–difficulty effects (specific, difficult vs. do-your-best or no goal). Full details of studies from 1966 to August 1984 are reported in Tables 1 and 2 in the Mento et al. (1987) article. The unpublished studies that were added to the Mento et al. studies were Chesney (1986), Shaw (1984), and Smith, Locke, and Barry (1986). Studies included correlational and experimental designs as well as laboratory and field settings. Moderator analyses conducted by Mento et al. showed that the size of effects for goal difficulty and goal specificity–difficulty were not affected by study design.

For experimental studies, results were converted to the effect-size statistic d . Results from correlational studies were first converted to point-biserial r (r_{pb}). These were then converted to an effect-size d to provide a common statistic for cumulating effect sizes across both correlational and experimental studies. Because the size of the point-biserial correlation is affected by the relative proportion of cases in the two treatment groups, effect sizes for all of the individual studies were corrected for differences in subgroup sample sizes when appropriate (Hunter et al., 1982). When cumulating results, effect sizes were weighted by the sample sizes for studies, as recommended by Hunter et al. (1982).

Of particular importance to the analyses and to our later discussion of results were the reliabilities of measures used for the performance criteria (in all studies) and the predictor variables (in field studies that

used a questionnaire measure of goal difficulty and goal specificity). The average reported reliabilities of criteria were .80 for studies using ratings of performance and .92 for studies with objective measures of performance. For the predictor variables in field studies, the average reliabilities were .72 (lowest .66) for goal difficulty and .81 (lowest .70) for goal specificity.

These reliability estimates were used to correct effect-size statistics and variance estimates (s_e^2) for error of measurement (cf. Hunter et al. 1982; Mabe & West, 1982). Next, sampling error variance was calculated using the formulas for sampling error modified to take into account the effect of the corrections for errors in measurement on sampling error, according to Hunter et al., 1982. Finally, the remaining unexplained variance (s_u^2) was determined after correcting for measurement error and sampling. For both the goal difficulty and the goal specificity–difficulty studies, the ratio of sampling error variance to total true variance (i.e., s_e^2) was less than the 75% cutoff suggested by Schmidt et al. (1982). Therefore, supplemental analyses to test for the moderator effects of task complexity were considered appropriate.

Task-complexity scores for each of the 125 studies included in the meta-analyses were obtained through ratings of the tasks used. Descriptions of the study tasks were given by one of the present authors to the other two, who independently rated each of them on a 10-point complexity scale on the basis of the general definition of task complexity in Wood (1986). Complexity involves three aspects: component complexity (number of acts and information cues involved), coordinative complexity (type and number of relationships among acts and cues), and dynamic complexity (changes in acts and cues and the relationships among them). These were used to code the tasks used in goal-setting studies on a common scale of complexity.

The correlation between the two independent sets of ratings was .92. For initial ratings that differed by 2 or more points on the 10-point scale, the raters discussed their differences and reached a consensus on the appropriate rating. This then became the complexity score that was used in the moderator analyses. For all other tasks, the complexity scores used in the analyses were the averages of two ratings. Examples of the complexity scores (rounded) assigned to different tasks from goal-setting studies are shown in Figure 1. Note that the highest rating given on the 10-point scale was 7. The task-complexity scores were entered as the moderator variable into a regression analysis, using the corrected study statistic (i.e., d) as the criterion variable.

Results

Tasks that have been used in goal-setting studies show a strong bias toward more simple tasks such as brainstorming, perceptual speed, and toy-assembly tasks. The frequency distribution of Goal-Setting Studies \times Task Complexity (Figure 2) is clearly skewed to the right, with most being at the less complex end of the scale. However, among the large number of studies that have been conducted, significant numbers of studies have

used more complex tasks. It was felt that there were enough studies at different levels across the range of task complexity to provide an adequate test of the moderator hypotheses.

For the 72 studies of goal difficulty ($N = 7,548$), the mean effect size, corrected for measurement error, was $d = .5770$, and the variance corrected for measurement error was $s_e^2 = .1487$. The unexplained variance after correcting for sampling error was .1014; therefore, only 32% of the variance in results across the studies was attributable to sampling error. In the subset of 37 studies ($n = 3,377$) with experimental designs, the corrected d was .6171. The associated variance, corrected for measurement error was $s_e^2 = .2118$, and 26% of this was attributable to sampling error.

Across the 53 studies of goal specificity–difficulty ($N = 6,635$), the mean corrected effect size was $d = .4305$, and the variance corrected for measurement error was $s_e^2 = .0626$. The variance due to sampling error was .0374, or 60% of the corrected variance. For the subset of 44 experimental studies ($n = 4,722$), the mean corrected d was .4669 and the associated $s_e^2 = .0732$, 60% of which was attributable to sampling error.

Clearly, both goal difficulty and goal specificity have significant relationships with performance, and there is sufficient unexplained variance in the strength of these relationships across studies to warrant investigation of potential moderators.

Moderator Analyses

The results of the regression analyses (Table 1) support both hypotheses. Task complexity was a significant moderator of the size of the performance relationship for goal difficulty and goal specificity–difficulty. Both betas were negative, indicating that the magnitude of goal to performance effects *decreased* as task complexity increased.

Table 2 shows the average corrected effect-size d s for a subgrouping of studies. The magnitude of effects for both goal specificity–difficulty and goal difficulty are more pronounced on simple tasks than on complex tasks. This result was consistent across many different subgroupings of studies (not shown), indicating that the result was not an artifact of the groupings used. The finer the distinctions between subgroupings on the task complexity scale, the greater the differences between the simplest and most complex tasks for both goal difficulty and goal specificity–difficulty. Figure 3 shows the plot of the mean corrected d s for the different levels of task complexity. The moderating effects of complexity were more pronounced for the goal-difficulty–performance relationship than for the goal-specificity–difficulty–performance relationship.

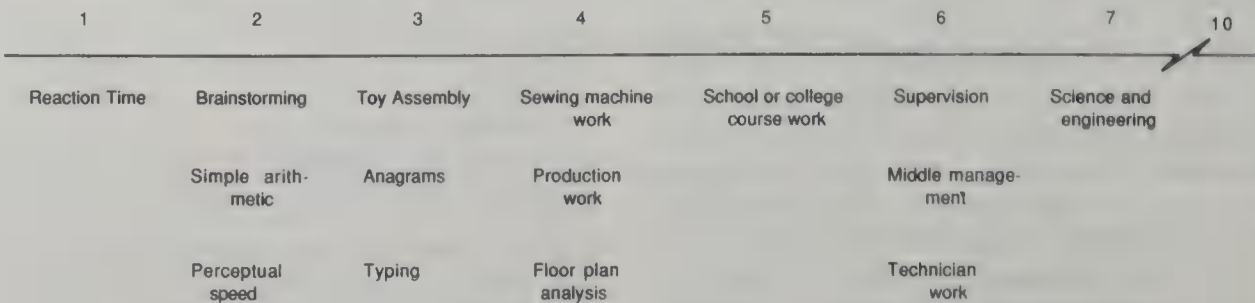


Figure 1. Representative tasks for various complexity levels.

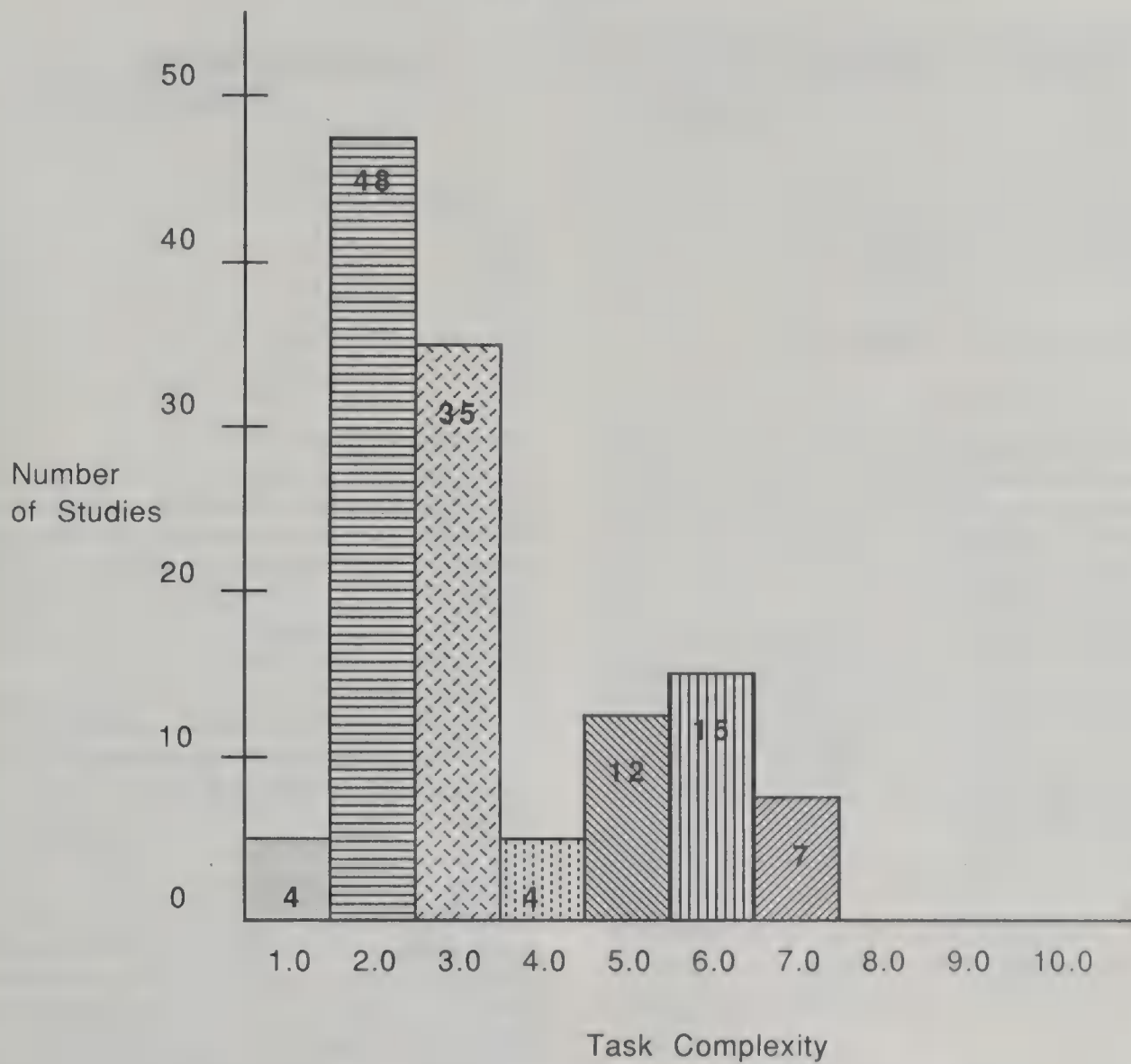


Figure 2. Frequency distribution of goal-setting studies by task complexity ($n = 125$).

It is interesting that across the full range of studies included in the regression analyses, the effects were stronger for goal specificity–difficulty ($R^2 = .0926$) than for goal difficulty ($R^2 = .0579$). This is the opposite of what would be expected after an inspection of Table 2 and Figure 3, and can be attributed to the difference in consistency of results in each type of study. The range of effect sizes was much greater in the goal-difficulty studies.

In Table 3, the d s for all studies combined, subgrouped into

quintiles by numbers of studies, are shown. The plot of these results (Figure 4) shows that, overall, the moderating effects on the relationships between specific, difficult goals and performance are most pronounced at the lowest levels of complexity. For studies in the middle range of the task-complexity scale (2.25–5.00), there was little difference in the goal–performance relationship.

Criterion Reliabilities

One possible explanation for our results is that performance measures for complex tasks, because they often involve less standardized outputs, are less reliable than performance measures for simple tasks. Therefore, it could be argued that the differences in goal effects between complex and simple tasks are an artifact of the differences in criterion reliabilities. This argument does not apply to the current results because the size of effects for each study were adjusted for the reported reliability of the criterion used.

However, to further test the potential validity of the argument, sensitivity analyses were conducted to see how great the differences in criterion reliabilities had to be in order to remove

Table 1
Multiple Regression Analysis Using Task Complexity as a Potential Moderator to Predict Goal Difficulty and Goal Specificity–Difficulty Performance Effects

Variable	No. of studies	β	R^2
Goal difficulty	72	-.240	.0579*
Goal specificity–difficulty	53	-.304	.0926*

* $p < .05$.

Table 2
Meta-Analysis ds for Task Complexity Subgroupings for Goal Difficulty and Goal Specificity–Difficulty

Task complexity rating	Goal difficulty			Goal specificity–difficulty		
	<i>n</i>	No. of studies	<i>ds</i> ^a	<i>n</i>	No. of studies	<i>ds</i>
1.0–2.75	3,297	33	.6941	2,115	19	.4727
3.0–4.75	1,615	18	.4991	3,138	21	.4338
5.0–7.25	2,636	21	.4781	1,382	13	.3583

^a *ds* are meta-analytic corrected for measurement error.

the observed effects. The 125 studies were split into two groups representing simple and complex tasks. The subgroup analyses were then rerun, assigning all studies in the simple group a reliability of 1 and all studies in the complex group a range of differing reliabilities (.8, .7, .6, .5, etc.). The results of these analyses showed that the differences between the two groups disappeared when the criterion reliability for the complex group was dropped to just below .6. Therefore, it appears that the observed differences in goal effects are sensitive to differences in the reliabilities of performance measures of .4 and above. However, for the current set of studies, differences in the criterion reliabilities between complex and simple tasks were not the cause of the differences in the size of goal effects.

Discussion

The results of our analyses provide strong evidence for the hypothesized moderating effects of task complexity on the relationships between the goal attributes and task performance.

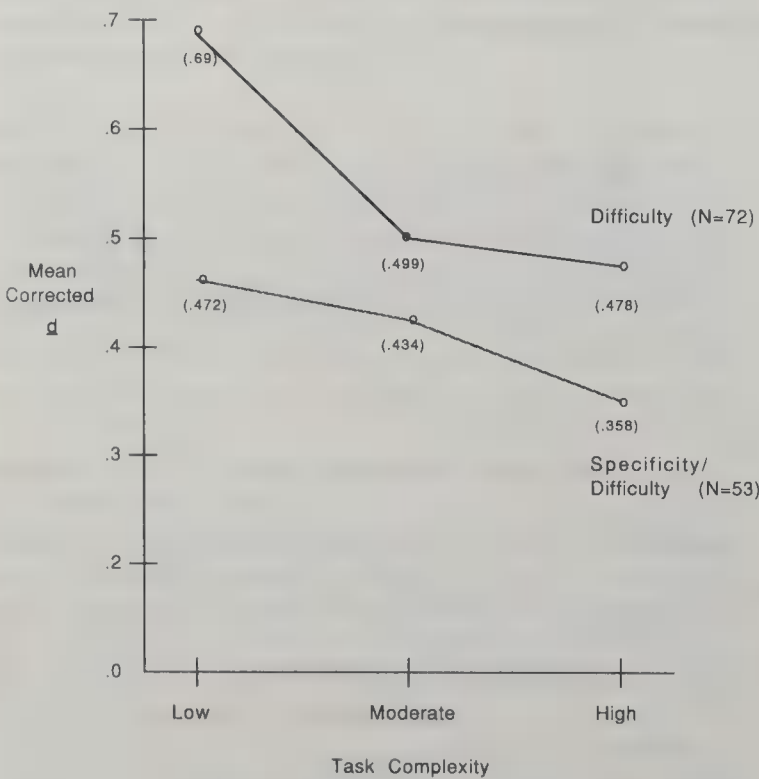


Figure 3. Goal effect as a function of task complexity (separately by each set).

The magnitude of goal effects on performance was greater on simple tasks than on complex tasks, and these results were not an artifact of differences in the reliability of the criterion measures. Therefore, to this point, task complexity is the only variable that has been shown to have a significant and robust moderating effect on the performance gains that result from specific, difficult goals. Mento et al., (1987) found inconsistent evidence of a moderator effect for feedback. (This may have been due to the fact that most of the studies in which feedback was manipulated could not be included in their moderator analysis.)

This finding complements the theoretical discussion and evidence for the strategy development effects of goals presented by Wood, Locke, and Smith (1986) and others (e.g., Campbell, 1984; Locke et al., 1981) who have begun to examine the processes by which goals affect performance on complex tasks. Of particular relevance to this study is the Wood et al. (1986) model, which predicts that the performance effects of goals will be lagged on complex tasks. This was based on studies by Shaw (1984) and Smith et al. (1986), who found that specific challenging goals lead to significant performance effects—but only in later trials—of a complex clerical task and a complex decision game, respectively. It is possible that differences in goal effects between complex and simple tasks may disappear over repeated performances of a task, as individuals develop effective strategies for the performance of complex tasks. Therefore, future research into the effects of goals on complex tasks should use multitrial or longitudinal methods that allow for the development of lagged effects (e.g., Campbell, 1984; Ivancevich, 1976; Shaw, 1984; Smith et al., 1986).

The benefits of more rigorous, standardized definitions of task characteristics, such as those advanced by Wood (1986), need to be recognized by goal researchers. The earlier review of goal-setting studies that examined task effects clearly demonstrated the problems associated with inadequate conceptualizations of and hypotheses about moderator effects. The coding of study tasks for the meta-analysis has allowed us to examine the previously uncontrolled sources of variation in goal-setting studies that were due to the task characteristic of complexity. Research is now needed to focus on how goal effects vary as a function of different types of complexity and the underlying processes by which goals affect performance on different types of tasks.

For example, we could consider *how* goals affect performance on complex psychomotor tasks. Much of the discussion about the effects of goals on complex tasks is focused on cognitive

Table 3
Meta-Analysis d s for Task Complexity Subgroups, by Quintiles, for All Studies Combined^a

Task complexity rating	<i>n</i>	No. of studies	d^b
1.0–2.0	2,702	27	.7672
2.25–2.75	2,710	25	.4485
3.0–3.25	2,082	20	.4370
3.5–5.0	4,234	28	.4697
5.25–7.0	2,455	25	.4173

^a Effects of goal specificity–difficulty and goal difficulty are combined.

^b d s are meta-analytic corrected for measurement error.

processes, such as search, information processing, and strategy development, which are typically required in the performance of complex decision tasks (e.g., Campbell, 1984; Wood et al., 1986). However, these cognitive processes may have less influence on the motor aspects of tasks in which performance requires a highly programmed set of behaviors as well as goals that focus the person's attention on the outcomes to be achieved. In fact, without a set of well-learned and effective motor programs, the attentional demands of outcome goals may undermine a person's execution of complex motor activities. Therefore, although the results would fit with our moderator hypothesis (i.e., outcome goals would lead to lower performance on more complex tasks), the underlying processes by which goals affect performance on complex psychomotor tasks may be quite different from those for more cognitive types of tasks.

One final implication of our results concerns the validity generalization arguments that Schmidt and Hunter (1977) have developed in relation to selection processes, but more recently have applied to other organization interventions, including goal setting (Hunter & Schmidt, 1983). There are two aspects to the Schmidt and Hunter arguments. The first is the validity generalization thesis that observed effects (validity of selection tests, performance effects of goals, etc.) generalize across a variety of organizations, jobs, and tasks. The results reported here, and in Mento et al. (1987), clearly support the thesis that the positive performance effects of goal difficulty–specificity are highly generalizable.

The second aspect of the Schmidt and Hunter (1977) arguments relates to the situational specificity hypothesis. That is, differences in situations will be associated with differences in the magnitude of effects. There is evidence that task complexity may be an important moderator across a range of performance determinants. For example, Hunter (1983) found that job complexity affected the selection validities for tests of cognitive abilities, with validities increasing with job complexity. This finding has been replicated by Gutenberg, Arvey, Osburn, and Jeaneret (1983), whose measure of job complexity was based on selected dimensions from the Position Analysis Questionnaire (PAQ). In both the Hunter and Gutenberg et al. studies, complexity was defined as the level of information-processing and problem-solving demands of the job. These task demands are products of the coordinative and dynamic types of task complexity defined by Wood (1986) and used to classify tasks in the present study.

In another area of selection research, tentative support has been found for the hypothesis that task complexity moderates the effectiveness of realistic job previews (RJPs) in reducing turnover (McEvoy & Cascio, 1985). The RJPs were found to be less effective in reducing turnover for entry-level nonmanagement jobs of low complexity than for more complex jobs. However, this result was based on a small number of studies for each level of task complexity, and the possibility of capitalization on chance could not be ruled out (McEvoy & Cascio, 1985, p. 349).

There is an interesting complementarity between our findings and those in the selection research area. Tests of cognitive abilities and RJPs are most effective in selecting for complex jobs, in which goal setting is least effective, and least effective for less complex jobs, in which goal setting is most effective. There is evidence that the effects of goal setting are mediated by strategy development—information processing and problem solving—on complex tasks (Wood et al., 1986). In the future, researchers need to consider the relative effects of cognitive abilities, information sharing (as in RJPs and participation) and goal setting, on strategy development.

Our support for the situational specificity hypothesis has implications for generalizing about the size of effects or productivity gains from goal setting across different types of tasks. The average effect-size goals for all of the studies combined was $d = .521$, equivalent to a 10.39% increase in productivity (Mento et al., 1987). For tasks coded at the low, moderate, and high levels of complexity, the equivalent productivity increases are 12.15%, 9.12%, and 7.79%, respectively.

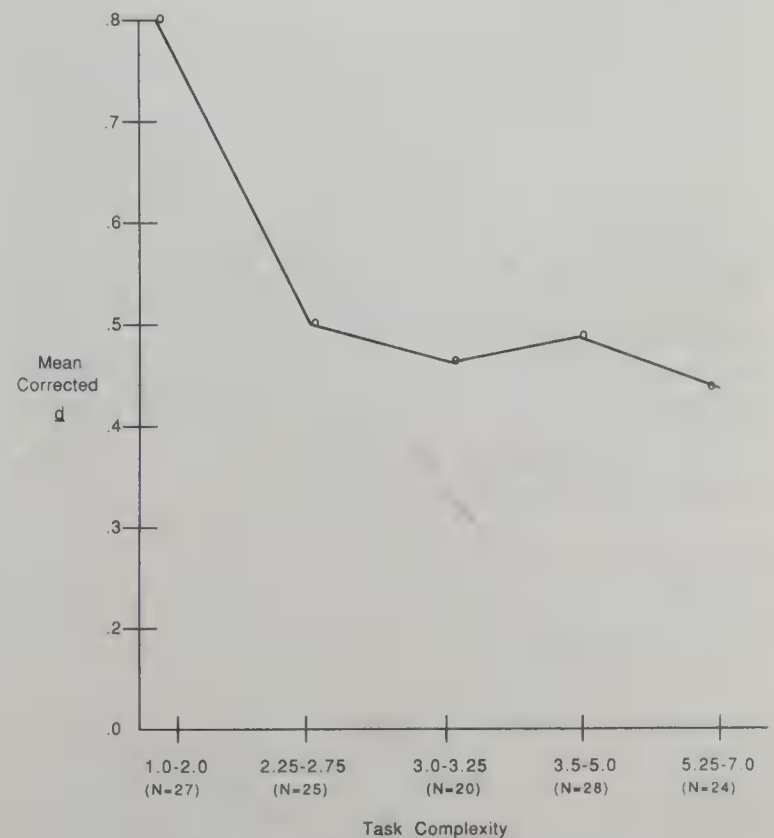


Figure 4. Goal effect as a function of task complexity (combined studies) by quintiles.

References

- Austin, J. T., & Bobko, P. (1984). The application of goal setting: Some boundary conditions and future research. *Academy of Management Proceedings*, 197-201.
- Baumler, J. V. (1971). Defined criteria of performance in organizational control. *Administrative Science Quarterly*, 16, 340-348.
- Campbell, D. J. (1984). The effect of goal contingent payment on the performance of a complex task. *Personnel Psychology*, 37, 23-40.
- Chesney, A. (1986). *An examination of the relation among goals, strategies, and performance: A simulation study*. Unpublished doctoral dissertation, University of Maryland.
- Fleishman, E. A. (1975). Toward a taxonomy of human performance. *American Psychologist*, 30, 1127-1149.
- Frost, P. J., & Mahoney, T. A. (1976). Goal setting and the task process: 1. An interactive influence on individual performance. *Organizational Behavior and Human Performance*, 17, 328-350.
- Green, B. F., & Hall, J. A. (1984). Quantitative methods for literature reviews. *Annual Review of Psychology*, 35, 37-53.
- Gutengberg, R. L., Arvey, R. D., Osburn, H. G., & Jeanneret, P. R. (1983). Moderating effects of decision-making/information-processing job dimensions on test validities. *Journal of Applied Psychology*, 68, 602-608.
- Hackman, J. R. (1969). Toward understanding the role of tasks in behavioral research. *Acta Psychologica*, 31, 97-128.
- Hackman, J. R., & Oldham, G. R. (1980). *Work redesign*. Reading, MA: Addison-Wesley.
- Huber, V. L. (1985). Effects of task difficulty, goal setting, and strategy on performance of a heuristic task. *Journal of Applied Psychology*, 70, 492-504.
- Hunter, J. E. (1983). *Test validation for 12,000 jobs: An application of job classification and validity generalization to the General Aptitude Test Battery (GATB)*. Washington, DC: U.S. Department of Labor, Employment and Training Administration, Division of Counseling and Test Development.
- Hunter, J. E., & Schmidt, F. L. (1983). Quantifying the effects of psychological interventions on employee job performance and work-force productivity. *American Psychologist*, 38, 473-478.
- Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.
- Ivancevich, J. M. (1976). Effects of goal setting on performance and job satisfaction. *Journal of Applied Psychology*, 61, 605-612.
- Jackson, S. E., & Zedeck, S. (1982). Explaining performance variability: Contributions of goal setting, task characteristics, and evaluative contexts. *Journal of Applied Psychology*, 67, 759-768.
- Latham, G. P., Mitchell, T. R., & Dossett, D. L. (1978). The importance of participative goal setting and anticipated rewards on goal difficulty and job performance. *Journal of Applied Psychology*, 63, 163-171.
- Locke, E. A. (1967). Relationship of goal level to performance level. *Psychological Reports*, 20, 1068.
- Locke, E. A., & Bryan, J. F. (1968). Grade goals as determinants of academic achievement. *Journal of General Psychology*, 79, 217-228.
- Locke, E. A., Shaw, K. N., Saari, L. M., & Latham, G. P. (1981). Goal setting and task performance: 1969-1980. *Psychological Bulletin*, 90, 125-152.
- Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology*, 67, 280-296.
- McEvoy, G. M., & Cascio, W. F. (1985). Strategies for reducing employee turnover: A meta-analysis. *Journal of Applied Psychology*, 70, 342-353.
- Mento, A. J., Steel, R. P., & Karren, R. J. (1987). A meta-analytic study of the effects of goal setting on task performance: 1966-1984. *Organizational Behavior and Human Decision Processes*, 39, 52-83.
- Naylor, J. C., & Ilgen, D. R. (1984). Goal-setting: A theoretical analysis of a motivational technology. In B. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. 6, pp. 95-140). Greenwich, CT: JAI Press.
- Peterson, N. G., & Bownas, D. A. (1982). Skill, task structure and performance acquisition. In M. D. Dunnette & E. A. Fleishman (Eds.), *Human performance and productivity* (Vol. 1, pp. 49-105). Hillsdale, NJ: Erlbaum.
- Pinder, C. (1984). *Work motivation*. Glenview, IL: Scott, Foresman.
- Rosenthal, R. (1978). Combining results from independent studies. *Psychological Bulletin*, 85, 185-193.
- Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, 62, 529-540.
- Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1982). Progress in validity generalization: Comments on Callender and Osborn and further developments. *Journal of Applied Psychology*, 67, 835-845.
- Shaw, K. N. (1984). *A laboratory investigation of the relationships among goals, strategies, and task performance*. Unpublished doctoral dissertation, University of Maryland.
- Smith, K. G., Locke, E. A., & Barry, D. (1986). *Goal setting, planning and organizational performance: An experimental study*. Unpublished manuscript, University of Maryland.
- Steel, R. P., & Ovalle, N. K. (1984). A review and meta-analysis of research on the relationship between behavioral intentions and employee turnover. *Journal of Applied Psychology*, 69, 673-686.
- Streufert, S., & Streufert, S. (1978). *Behavior in the complex environment*. New York: Wiley.
- Weick, K. E. (1965). Laboratory experimentation with organizations. In J. G. March (Ed.), *Handbook of organizations*. Chicago: Rand-McNally.
- Wood, R. E. (1985, August). *Task complexity and goal effects*. Paper presented at the meeting of the National Academy of Management, San Diego, CA.
- Wood, R. E. (1986). Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes*, 37, 60-82.
- Wood, R. E., Locke, E. A., & Smith, K. (1986). *Goal setting and strategy effects on complex tasks: A theoretical analysis*. Unpublished manuscript, University of New South Wales, New South Wales, Australia.

Appendix

References for the Meta Analysis

The letters at the end of each reference mean the following: Correlational studies are designated *C* and experimental articles *E*. Those articles involving goal difficulty have a *D* and those comparing specific, difficult goals with do-best goals are designated *SD*.

- Andrews, F. M., & Farris, G. F. (1972). Time pressure and performance of scientists and engineers: A five year panel study. *Organizational Behavior and Human Performance*, 8, 185–200. C, D
- Bandura, A., & Cervone, D. (1983). Self-evaluative and self-efficacy mechanisms governing the motivational effects of goal systems. *Journal of Personality and Social Psychology*, 45, 1017–1028. SD, E
- Bandura, A., & Schunk, D. H. (1981). Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. *Journal of Personality and Social Psychology*, 41, 586–598. SD, E
- Basset, G. A. (1979). A study of the effects of task goal and schedule choice on work performance. *Organizational Behavior and Human Performance*, 24, 202–227. D, E
- Bavelas, J. B., & Lee, E. S. (1978). Effects of goal level on performance: A tradeoff of quantity and quality. *Canadian Journal of Psychology*, 32, 219–240. D, E
- Becker, L. J. (1978). Joint effect of feedback and goal setting on performance: A field study of residential energy conservation. *Journal of Applied Psychology*, 63, 428–433. D, E
- Blumenfeld, W. S., & Leidy, T. R. (1969). Effectiveness of goal setting as a management device: Research note. *Psychological Reports*, 24, 752. SD, C
- Burke, R. J., & Wilcox, D. S. (1969). Characteristics of effective employee performance review and development. *Personnel Psychology*, 22, 291–305. SD, C
- Campbell, D. J. (1984). The effect of goal contingent payment on the performance of a complex task. *Personnel Psychology*, 37, 23–40. D, E
- Campbell, D. J., & Ilgen, D. R. (1976). Additive effects of task difficulty and goal setting on subsequent task performance. *Journal of Applied Psychology*, 61, 319–324. D, E
- Chesney, A. (1986). *An examination of the relation among goals, strategies, and performance: A simulation study*. Unpublished doctoral dissertation, University of Maryland. D, E
- Dachler, H. P., & Mobley, W. H. (1973). Construct validation of an instrumentality-expectancy-task-goal model of work motivation: Some theoretical boundary conditions [Monograph]. *Journal of Applied Psychology*, 58, 397–418. D, C
- Dossett, D. L., Latham, G. P., & Mitchell, T. R. (1979). The effects of assigned versus participatively set goals, KR, and individual differences when goal difficulty is held constant. *Journal of Applied Psychology*, 64, 291–298. D, C
- Dossett, D. L., Latham, G. P., & Saari, L. M. (1980). The impact of goal setting on survey returns. *Academy of Management Journal*, 23, 561–567. SD, E
- Erez, M. (1977). Feedback: A necessary condition for the goal setting-performance relationship. *Journal of Applied Psychology*, 62, 624–627. D, C
- Frost, P. J., & Mahoney, T. A. (1976). Goal setting and the task process: 1. An interactive influence on individual performance. *Organizational Behavior and Human Performance*, 17, 328–350. SD, E
- Garland, H. (1982). Goal levels and task performance: A compelling replication of some compelling results. *Journal of Applied Psychology*, 67, 245–248. D, E
- Garland, H. (1983). Influence of ability, assigned goals, and normative information on personal goals and performance: A challenge to the goal attainability assumption. *Journal of Applied Psychology*, 68, 20–30. D, E
- Garland, H. (1984a, August). *A cognitive mediation theory of task goals and human performance*. Paper presented at the 44th Annual Convention of the Academy of Management, Boston. D, E
- Garland, H. (1984b). Relation of effort-performance expectancy to performance in goal-setting experiments. *Journal of Applied Psychology*, 69, 79–84. SD, E
- Hall, D. T., & Foster, L. W. (1977). A psychological success cycle and goal setting: Goals, performance, and attitudes. *Academy of Management Journal*, 20, 282–290. D, C
- Hall, D. T., & Hall, F. S. (1976). The relationship between goals, performance, success, self-image, and involvement under different organizational climates. *Journal of Vocational Behavior*, 9, 267–278. D, C
- Hamner, W. C., & Harnett, D. L. (1974). Goal-setting, performance and satisfaction in an interdependent task. *Organizational Behavior and Human Performance*, 12, 217–230. D, E
- Ivancevich, J. M. (1976). Effects of goal setting on performance and job satisfaction. *Journal of Applied Psychology*, 61, 605–612. SD, E
- Ivancevich, J. M. (1977). Different goal-setting treatments and their effects on performance and job satisfaction. *Academy of Management Journal*, 20, 406–419. SD, C
- Ivancevich, J. M. (1982). Subordinates' reactions to performance appraisal interviews: A test of feedback and goal-setting techniques. *Journal of Applied Psychology*, 67, 581–587. SD, E
- Ivancevich, J. M., & McMahon, J. T. (1977a). Black-white differences in a goal-setting program. *Organizational Behavior and Human Performance*, 20, 287–300. D, C; SD, C
- Ivancevich, J. M., & McMahon, J. T. (1977b). Education as a moderator of goal-setting effectiveness. *Journal of Vocational Behavior*, 11, 83–94. D, C; SD, C
- Ivancevich, J. M., & McMahon, J. T. (1977c). A study of task-goal attributes, higher order need strength, and performance. *Academy of Management Journal*, 20, 552–563. D, C; SD, C
- Ivancevich, J. M., & McMahon, J. T. (1982). The effects of goal setting, external feedback, and self-generated feedback on outcome variables: A field experiment. *Academy of Management Journal*, 25, 359–372. SD, E
- Jackson, S. E., & Zedeck, S. (1982). Explaining performance variability: Contributions of goal setting, task characteristics, and evaluative contexts. *Journal of Applied Psychology*, 67, 759–768. D, E
- Kaplan, R., & Rothkopf, E. Z. (1974). Instructional objectives as directions to learners: Effect of passage length and amount of objective-relevant content. *Journal of Educational Psychology*, 66, 448–456. SD, E
- Kim, J. S. (1984). Effect of behavior plus outcome goal setting and feedback on employee satisfaction and performance. *Academy of Management Journal*, 27, 139–149. SD, E
- LaPorte, R. E., & Nath, R. (1976). Role of performance goals in prose learning. *Journal of Educational Psychology*, 68, 260–264. D, E
- Latham, G. P., & Kinne, S. B. III. (1974). Improving job performance through training in goal setting. *Journal of Applied Psychology*, 59, 187–191. SD, E
- Latham, G. P., & Locke, E. A. (1975). Increasing productivity with decreasing time limits: A field replication of Parkinson's Law. *Journal of Applied Psychology*, 60, 524–526. SD, E
- Latham, G. P., & Marshall, H. A. (1982). The effects of self-set, participatively set, and assigned goals on the performance of government employees. *Personnel Psychology*, 35, 399–404. D, C
- Latham, G. P., Mitchell, T. R., & Dossett, D. L. (1978). The importance

- of participative goal setting and anticipated rewards on goal difficulty and job performance. *Journal of Applied Psychology*, 63, 163-171. D, C
- Latham, G. P., & Saari, L. M. (1979a). The effects of holding goal difficulty constant on assigned and participatively set goals. *Academy of Management Journal*, 22, 163-168. SD, E
- Latham, G. P., & Saari, L. M. (1979b). The importance of supportive relationships in goal setting. *Journal of Applied Psychology*, 64, 151-156. D, C; SD, E
- Latham, G. P., & Steel, T. P. (1983). The motivational effects of participative versus assigned goal setting on performance. *Academy of Management Journal*, 26, 406-417. D, C
- Latham, G. P., & Yukl, G. A. (1976). The effects of assigned and participative goal setting on performance and job satisfaction. *Journal of Applied Psychology*, 61, 166-171. SD, E
- Locke, E. A. (1966). The relationship of intentions to level of performance. *Journal of Applied Psychology*, 50, 60-66. D, E
- Locke, E. A. (1967). The motivational effects of knowledge of results: Knowledge or goal-setting? *Journal of Applied Psychology*, 51, 324-329. SD, E
- Locke, E. A. (1968). The effects of knowledge of results, feedback in relation to standards, and goals on reaction time performance. *American Journal of Psychology*, 81, 566-574. D, E
- Locke, E. A. (1982). Relation of goal level to performance with a short work period and multiple goal levels. *Journal of Applied Psychology*, 67, 512-514. D, E
- Locke, E. A., & Bryan, J. F. (1966). Cognitive aspects of psychomotor performance: The effects of performance goals on level of performance. *Journal of Applied Psychology*, 50, 417-420. SD, E
- Locke, E. A., & Bryan, J. F. (1967). Performance goals as determinants of level of performance and boredom. *Journal of Applied Psychology*, 51, 120-130. D, C; SD, E
- Locke, E. A., & Bryan, J. F. (1968). Grade goals as determinants of academic achievement. *Journal of General Psychology*, 79, 217-228. D, C
- Locke, E. A., & Bryan, J. F. (1969). The directing function of goals in task performance. *Organizational Behavior and Human Performance*, 4, 35-42. D, E; SD, E
- Locke, E. A., Bryan, J. F., & Kendall, L. M. (1968). Goals and intentions as mediators of the effects of monetary incentives on behavior. *Journal of Applied Psychology*, 52, 104-121. D, C
- Locke, E. A., Cartledge, N., & Kner, C. (1970). Studies of the relationship between satisfaction, goal setting and performance. *Organizational Behavior and Human Performance*, 5, 135-158. D, C
- Locke, E. A., Frederick, E., Lee, C., & Bobko, P. (1984). Effects of self-efficacy, goals, and task strategies on task performance. *Journal of Applied Psychology*, 69, 241-251. D, C
- Locke, E. A., Mento, A. J., & Katcher, B. (1978). The interaction of ability and motivation in performance: An exploration of the meaning of moderators. *Personnel Psychology*, 31, 269-280. SD, E
- Locke, E. A., & Shaw, K. N. (1984). Atkinson's inverse-U curve and the missing cognitive variables. *Psychological Reports*, 55, 403-412. D, C
- London, M., & Oldham, G. R. (1976). Effects of varying goal types and incentive systems on performance and satisfaction. *Academy of Management Journal*, 19, 537-546. D, E
- Masters, J. C., Furman, W., & Barden, R. C. (1977). Effects of achievement standards, tangible rewards and self-dispersed achievement evaluations on children's task mastery. *Child Development*, 48, 217-224. D, E
- Matsui, T., Okada, A., & Kakuyama, T. (1982). Influence of achievement need on goal setting, performance, and feedback effectiveness. *Journal of Applied Psychology*, 67, 645-648. D, C
- McCaul, K. D., & Kopp, J. T. (1982). Effects of goal setting and commitment on increasing metal recycling. *Journal of Applied Psychology*, 67, 377-379. SD, E
- Mento, A. J., Cartledge, N. D., & Locke, E. A. (1980). Maryland vs. Michigan vs. Minnesota: Another look at the relationship of expectancy and goal difficulty to task performance. *Organizational Behavior and Human Performance*, 25, 419-440. D, E
- Mossholder, K. W. (1980). Effects of externally mediated goal setting on intrinsic motivation: A laboratory experiment. *Journal of Applied Psychology*, 65, 202-210. SD, E
- Motowidlo, S., Loehr, V., & Dunnette, M. D. (1978). A laboratory study of the effects of goal specificity on the relationship between probability of success and performance. *Journal of Applied Psychology*, 23, 172-179. D, E
- Mowen, J. C., Middlemist, R. D., & Luther, D. (1981). Joint effects of assigned goal level and incentive structure on task performance: A laboratory study. *Journal of Applied Psychology*, 66, 598-603. D, E
- Nemeroff, W. F., & Cosentino, J. (1979). Utilizing feedback and goal setting to increase performance appraisal interviewer skills of managers. *Academy of Management Journal*, 22, 566-576. SD, E
- Oldham, G. R. (1976). The motivational strategies used by supervisors: Relationships to effectiveness indicators. *Organizational Behavior and Human Performance*, 15, 66-86. D, C
- Organ, D. W. (1977). Intentional vs. arousal effects of goal setting. *Organizational Behavior and Human Performance*, 18, 377-389. D, E
- Peters, L. H., Chassie, M. B., Lindholm, H. R., O'Connor, E. J., & Kline, C. R. (1982). The joint influence of situational constraints and goal setting on performance and affective outcomes. *Journal of Management*, 8, 7-20. D, E
- Pritchard, R. D., & Curtis, M. I. (1973). The influence of goal setting and financial incentives on task performance. *Organizational Behavior and Human Performance*, 10, 175-183. D, E
- Rakestraw, T. L. Jr., & Weiss, H. M. (1981). The interaction of social influence and task experience on goals, performance, and performance satisfaction. *Organizational Behavior and Human Performance*, 27, 326-344. D, C
- Ronan, W. W., Latham, G. P., & Kinne, S. B. (1973). Effects of goal setting and supervision on worker behavior in an industrial situation. *Journal of Applied Psychology*, 58, 302-307. SD, C
- Rosswork, S. G. (1977). Goal setting: The effects on an academic task with varying magnitudes of incentive. *Journal of Educational Psychology*, 69, 710-715. SD, E
- Rothkopf, E. Z., & Billington, M. J. (1975). A two-factor model of the effect of goal-descriptive directions on learning from text. *Journal of Educational Psychology*, 67, 192-204. D, E; SD, E
- Rothkopf, E. Z., & Billington, M. J. (1979). Goal-guided learning from text: Inferring a descriptive processing model from inspection times and eye movements. *Journal of Educational Psychology*, 71, 310-327. SD, E
- Rothkopf, E. Z., & Kaplan, R. (1972). Exploration of the effect of density and specificity of instructional objectives on learning from text. *Journal of Educational Psychology*, 63, 295-302. SD, E
- Sales, S. M. (1970). Some effects of role overload and role underload. *Organizational Behavior and Human Performance*, 5, 592-608. D, E
- Shaw, K. N. (1984). *A laboratory investigation of the relationships among goals, strategies, and task performance*. Unpublished doctoral dissertation, University of Maryland. D, E
- Smith, K. G., Locke, E. A., & Barry, D. (1986). *Goal setting, planning and organizational performance: An experimental study*. Unpublished manuscript, University of Maryland. D, E
- Steers, R. M. (1975). Task-goal attributes, achievement, and supervisory performance. *Organizational Behavior and Human Performance*, 13, 392-403. D, C; SD, C
- Strang, H. R., Lawrence, E. C., & Fowler, P. C. (1978). Effects of as-

- signed goal level and knowledge of results on arithmetic computation: A laboratory study. *Journal of Applied Psychology*, 63, 29–39. D, E
- Taylor, M. S., Locke, E. A., Lee, C., & Gist, M. (1984). Type A behavior and faculty research productivity: What are the mechanisms? *Organizational Behavior and Human Performance*, 34, 402–418. D, C
- Terborg, J. R. (1976). The motivational components of goal setting. *Journal of Applied Psychology*, 61, 613–621. D, C
- Terborg, J. R., & Miller, H. E. (1978). Motivation, behavior and performance: A closer examination of goal setting and monetary incentives. *Journal of Applied Psychology*, 63, 29–39. SD, E
- Umstot, D. D., Bell, C. H., Jr., & Mitchell, T. R. (1976). Effects of job enrichment and task goals on satisfaction and productivity: Implications for job design. *Journal of Applied Psychology*, 61, 379–394. SD, E
- Wexley, K. N., & Nemeroff, W. F. (1975). Effectiveness of positive reinforcement and goal setting as methods of management development. *Journal of Applied Psychology*, 60, 446–450. SD, E
- White, S. E., Mitchell, T. R., & Bell, C. H., Jr. (1977). Goal setting evaluation apprehension, and social cues as determinants of job performance and job satisfaction in a simulated organization. *Journal of Applied Psychology*, 62, 665–673. SD, E
- Wofford, J. C. (1982). Experimental tests of the goal–energy–effort requirement theory of work motivation. *Psychological Reports*, 50, 1259–1273. SD, E
- Wood, R. E., & Locke, E. A. (1984). *The effects of self-efficacy on academic performance*. Unpublished manuscript, University of New South Wales, New South Wales, Australia. D, C
- Yukl, G. A., & Latham, G. P. (1978). Interrelationships among employee participation, individual differences, goal difficulty, goal acceptance, goal instrumentality and performance. *Personnel Psychology*, 31, 305–323. D, C

Received July 14, 1986

Revision received February 5, 1987

Accepted December 12, 1986 ■

Differences Among Differences: In Search of General Work Preference Dimensions

Robert G. L. Pryor

New South Wales Department of Industrial Relations and Employment
Darlinghurst, New South Wales, Australia

An adequate investigation of differences between individuals, it is argued, would take account of differences across dimensions as well as along dimensions. Previously, researchers have not been able to provide adequate means to assess general dimensions of values and preferences related to work. In this study, the Work Aspect Preference Scale was administered to samples of Grade 10 students, Grade 11/12 students, and adults. Factor analyses of data from each sample were performed, yielding three similar general work preference dimensions across the samples. Scoring procedures were developed that served as a basis for concurrent validation research and test-retest reliability studies. Results suggested that the three factors—Non-Work Orientation, Human/Personal Concern, and Freedom—could be assessed conveniently, reliably, and validly. Comparison of the dimensions across the samples indicated the consistency of the three-dimension solution; however, some variations in the content of the dimensions were noted. Some applications of the derived dimensions are outlined.

Historically, the ways to study individual differences have ranged from the clinical and idiographic approaches of existential psychologists at one extreme to the nomothetic methods of psychometricians at the other. In this latter tradition, Guilford (1954, 1959, 1975) consistently called on researchers to consider individual differences from a broader perspective. He observed that most psychometric approaches to individual differences focused on differences *along* a dimension or set of dimensions. However, Guilford noted, there is no reason to assume that individuals may not also differ *across* dimensions. The work of personality theorists from the psychometric tradition, like Cattell and Eysenck, illustrate the point. Cattell (1946) hypothesized 16 specific factors, and Eysenck investigated two (Eysenck & Eysenck, 1969), and later, three (Eysenck & Eysenck, 1975) general dimensions. Cattell's (1973) subsequent work on second-order factors revealed that some of Eysenck's dimensions could be derived from his original specific factors. A further psychometric illustration of Guilford's observation can be seen in hierarchical models of intelligence such as that of Burt and Vernon (e.g., Vernon, 1979), which comprised different levels of specific abilities as well as more global group factors and general ability or intelligence. In the other measurement area with a long psychometric tradition, vocational interests, similar developments are occurring. Thus, the Strong-Campbell Interest Inventory (Campbell & Hansen, 1981), in addition to assessing the 24 occupational scales, assesses Holland's (1973) six typological themes. Furthermore, Athanasou (1986) sought to generalize the Holland types into Prediger's (1976) two bipolar dimensions. However, despite a plethora of

concepts and instruments (Cook, Hepworth, Wall, & Warr, 1981) in the field of the measurement of values and preferences related to work, there appears to be a noticeable lack of psychometrically adequate attempts to integrate and assess these characteristics at different levels of generality.

Most attempts to investigate levels of generality of values and preferences related to work consist of multivariate analyses of existing inventories. O'Connor and Kinnane (1961) factor analyzed an early version of Super's Work Values Inventory, identifying five factors: Security-Economic-Material, Social-Artistic, Work Conditions and Associates, Achievement-Prestige, and Independence-Variety. Lofquist and Dawis (1978) reported factor analyses of 20 scales of the Minnesota Importance Questionnaire and found that a six-factor solution comprising Safety, Autonomy, Comfort, Altruism, Achievement/Accomplishment, and Self-Aggrandizement, best accounted for the results. Such studies have two major limitations. First, they apply factor analytic techniques to ipsative measures, thereby violating the independence assumption of these techniques, and second, they usually provide no way of accurately assessing the second-order dimensions found. Even in studies in which the first objection cannot be sustained (e.g., Pryor, 1980), the latter very often can. Furthermore, research such as that of Boulton (1980), using the normative version of the Work Values Inventory and providing procedures whereby subsequent measurement of the six derived factors (Stimulating Work, Interpersonal Satisfaction, Economic Security, Responsible Autonomy, Comfortable Existence, and Esthetic Concern) could be undertaken, has other limitations. The individuals tested are atypical, being 445 physically disabled counseling clients. Furthermore, no norms or validity data for the six factors are offered in his or any of the other studies cited.

Pryor (1982) outlined a conceptual framework for linking needs, values, preferences, work ethics, and orientations to work. He maintained (a) that all of these terms are related to *liking* or *preferring* between the person and work, (b) that such

The author gratefully acknowledges the contribution of two anonymous reviewers to the comparison of factors analysis in this article.

Correspondence concerning this article should be addressed to Robert G. L. Pryor, Central Planning and Research Unit, N.S.W. Department of Industrial Relations and Employment, P.O. Box 847, Darlinghurst 2010, New South Wales, Australia.

terms differ in the level of generality of these relations, (c) that at each level of generality, the relevant relation is a valid explanation, (d) that such terms are not simply redefinitions of one another, (e) that a person may stand in more than one of these relations to work at any one time, and (f) that such relations in no way define either the person or the work context. Using complete-link cluster analysis, Pryor (1982, 1983a) sought to illustrate how the relations between these terms might be empirically integrated. However, this research had two limitations. First, although both studies compared dendrograms from different subsamples, no quantitative procedure was used for such comparison. Second, no ways were specified for the measurement of the clusters.

The present research program is an attempt to remedy some of the shortcomings evident in much of the previous research into the measurement of values and preferences related to work at different levels of generality. The research aims of the studies reported here were first, to investigate general work preference dimensions derived from more specific work aspect preferences (Pryor, 1979); second, to specify how these dimensions can be accurately and parsimoniously measured; third, to indicate how consistent or generalizable these work preference dimensions are across different samples; and fourth, to provide both validity and reliability data on these assessable dimensions. Four research studies on different samples are reported. Studies 1 and 2 investigated the factor patterns underlying specific work aspect preferences of senior high school students (Grades 11/12) and adults, respectively. Procedures are outlined for the measurement of these factors, and the scores derived from these procedures are used to study the concurrent validity of each general work preference dimension. Study 3 examined the factor pattern underlying preference for aspects of work by high school students in Grade 10. As in the earlier studies, procedures are established for the measurement of these general work dimensions. The scores derived from these procedures are compared with scores from earlier studies that investigated how generalizable the work preference dimensions might be. Study 4 examined the short-term and medium-term test-retest reliability of the general dimensions of work preference.

Study 1

Method

Subjects. Subjects were 1,228 New South Wales Grades 11/12 high school students. Mean age was 16.28 years; 55% of the subjects were boys, and 45% were girls.

Measure. The subjects were administered the Work Aspect Preference Scale (WAPS; Pryor, 1983b), a 52-item test of dimensions people consider important in work. The conceptual basis of this inventory is outlined in Pryor (1979). Pryor (1981a) outlined the derivation of the WAPS subscales through the combined use of factor analysis and cluster analysis. The WAPS measures 13 subscales, briefly outlined in Table 1 and accompanied by estimates of their internal consistency and test-retest reliability. A variety of studies (Pryor, 1983b) indicate that all of the subscales significantly differentiate occupational choices of students and the occupations that people are working in or preparing for. (For further information about validity, see Pryor, 1981b, and Taylor & Pryor, 1986.)

Procedure. The 13 subscale scores of the WAPS for the sample were factor analyzed using a variety of factoring methods and rotational procedures. Because all of the methods presented the same general factor

patterns, the principal components solution with a varimax rotation was chosen to be used. This is the least complicated of factoring procedures and has the advantage of using all the test score variance. To summarize and score sample variance, principal components analysis is preferable because unlike principal factoring it does not partition part of the variance as *error* and then disregard this proportion of the variance in subsequent analysis. Inasmuch as stepwise multiple regression analysis was used to obtain the most parsimonious combination of WAPS subscales for the prediction of the derived second-order factor scores, having access to the total variance through the use of principal components analysis was of statistical advantage. To assist in the interpretation and designation of each factor, only subscales with absolute value loadings of .55 or above are considered. To simplify factor scoring, appropriate weights for each factor were obtained from stepwise multiple regression. Four criteria were adopted for inclusion of various weighted WAPS subscales in each regression equation. They are as follows: (a) There should not be more than four subscales. (b) The combined factor variance explained should be at least 75%. (c) The addition of an extra subscale to the equation should increase the variance explained by more than 5%. (d) The equation as a result of including ■ subscale in relation to the original factor pattern should be interpretable.¹

To obtain some estimate of validity of the scores derived from the regression equations, a multivariate analysis of variance (MANOVA) tested whether the derived general work preference scores significantly varied with students occupational choice. The 20 occupations were as follows: engineering, professional; engineering, technical; building and architectural sciences; farming; medical profession; nursing; social science; law; humanities; teaching, high school; teaching, primary school; administrative/commercial, professional level; administrative/commercial, nonprofessional level; creative arts; armed services; and data processing.

Results and Discussion

Factor pattern. Table 2 presents the results of the principal components analysis of the WAPS subscale scores. Using the criterion of an eigenvalue greater than 1.0, three factors were extracted and rotated, accounting for 57.8% of the total variance. The amount of variance unexplained is fairly high, and because in principal components analysis, variance explained is directly related to eigenvalues, the data contain a high number of other factors accounting for very small amounts of the variance. This inferred that a large amount of specific factor variance may be a direct consequence of the factorial derivation of the original WAPS subscales.

The highest loadings for the first factor, Non-Work Orientation, are for Detachment, Life Style, and Money. This factor is labeled Non-Work Orientation inasmuch as the dominant loadings suggest a concern that work should interfere as little as possible with the rest of one's life and, at the same time, provide monetary means with which to live it. This dimension is similar to the "non-job oriented" person of Dubin and Champoux (1975) and those individuals with an "instrumental" work orientation, as described by Goldthorpe, Lockwood, Beckhofer, and Platt (1968) and Beynon and Blackburn (1972).

¹ When final multiple equations for each factor were derived, these equations were used to generate a distribution of students' scores. To avoid fractions and negative numbers, all of the scores derived from the equations were multiplied by 100, and the lower limit of the scores was added to each score. This information is available from the author and may be found in Pryor (1986).

Table 1
Work Aspect Preference Subscales and Reliabilities

Subscale	Description	Reliabilities	
		Internal consistency ^a	Test-retest ^b
Independence	A concern for being free from imposed constraints in the work environment	.67	.69
CoWorkers	A concern for friendship and understanding from those with whom one works	.85	.78
Self-Development	A concern for developing and using one's skills and abilities	.85	.76
Creativity	A concern for developing something original through one's work	.83	.80
Money	A concern for obtaining large financial rewards from one's work	.82	.88
Life Style	A concern for the effect that employment may have on where and how one lives	.71	.74
Prestige	A concern for recognition and status in the eyes of others	.83	.81
Altruism	A concern for assisting others	.87	.83
Security	A concern for being able to maintain one's job	.94	.83
Management	A concern for organizing the work of others	.83	.78
Detachment	A concern for being able to separate work and its influence from other parts of one's life	.87	.76
Physical Activity	A concern for being physically active in one's work	.76	.83
Surroundings	A concern for the kind of physical environment in which one works	.66	.74

^a Alpha coefficients for tertiary students; *n* = 140.
^b Two-week intervening period; *n* = 117.

The second factor is characterized by high positive loadings on Creativity, Self-Development, Independence, and Management. It is called Freedom, indicating a personal commitment to work in which the person exercises prerogative and initiates rather than follows, innovates rather than implements. This dimension appears to be similar to dimensions of “independence/variety” (O'Connor & Kinnane, 1961), “self-expression” (Hen-

drix & Super, 1968), “autonomy” (Lofquist & Dawis, 1978), and “freedom” (Pryor, 1980).

The third factor is dominated by three high positive loadings on the WAPS subscales, Altruism, CoWorkers, and Security. This factor, People Concern, suggests a dimension, found also by O'Connor and Kinnane (1961), Lofquist and Dawis (1978), and Pryor (1980), with the additional emphasis on lack of risk.

The three-factor pattern found in Table 1 resembles the environment-people-self trichotomy suggested by Lofquist and Dawis (1978). In addition, the findings in this study are similar to Mortimer's (1975) three-factor solution: Extrinsic, Self-Expression, and People.

Scoring the factors. Equations from the stepwise multiple regression analysis are presented in the first section of Table 3. The *Rs* range from .87 to .93, suggesting a satisfactorily close correspondence between the scores derived from the equations and the factor scores.

Concurrent validity. Using scores derived from the regression equations in the middle part of Table 3, a MANOVA was performed using the three second-order dimension scores as the dependent variables and occupational choice category as the independent variable. The significant overall $F(57, 2714) = 7.03$, $p < 0.1$, indicates substantial differences in scores across categories. Univariate *F* tests were conducted for each preference dimension: nonwork orientation, $F(19, 905) = 5.62$, $p < .01$; freedom, $F(19, 905) = 5.47$, $p < .01$; and people concern, $F(19, 905) = 6.31$, $p < .01$. These data suggest that each of the three dimensions is capable of differentiating among students' occupational choices.

Study 2

Method

Subjects. Subjects were 1,356 adults who had left high school; 54% were men, and 46% were women. Forty-eight percent were 20 years old or younger, 34% between 21 and 30 years, and 18% more than 30.

Table 2
Principal Components Analysis With Varimax Rotation for Grades 11/12 Students' Work Aspect Preference Scale Scores

Work aspect preference	Rotated factor loadings		
	Non-work orientation	Freedom	People concern
Subscale			
Independence	.54	.59	-.08
CoWorkers	.31	.16	.71
Self-Development	-.03	.69	.40
Creativity	.03	.78	.11
Money	.71	.36	.02
Life Style	.74	.14	.14
Prestige	.39	.48	.32
Altruism	-.19	.27	.75
Security	.48	.07	.57
Management	.23	.58	.36
Detachment	.76	-.11	.14
Physical Activity	.15	.39	.41
Surroundings	.52	.19	.51
Eigenvalue	4.83	1.63	1.05
Percentage of common variance	64.30	21.70	14.00
Percentage of total variance	37.20	12.50	8.10

Note. *N* = 1,228.

Table 3
Multiple Regression Approximations to Factor Scores Across Three Samples

Factor	Predictors	R
Grade 11/12 students		
Non-Work Orientation	.11 (Det.) + .10 (Mon.) + .08 (L.S.) - 3.59	.93
Freedom	.12 (Cre.) + .10 (S.D.) + .08 (Ind.) - 4.09	.89
People Concern	.13 (Alt.) + .14 (CoW.) - 3.91	.87
Adults		
Human/Personal Concern	.15 (S.D.) + .08 (Sec.) + .09 (Alt.) - 4.98	.90
Non-Work Orientation	.12 (Det.) + .13 (Mon.) + .07 (L.S.) - 4.01	.93
Freedom	.18 (Cre.) + .17 (Ind.) - 4.70	.92
Grade 10 students		
Non-Work Orientation	.13 (L.S.) + .11 (Det.) + .11 (CoW.) - 5.20	.91
Power	.08 (Pre.) + .13 (Mon.) + .13 (Man.) - 4.52	.90
Human/Personal Concern	.22 (S.D.) + .13 (Alt.) - .08 (Mon.) - 4.20	.90

Note. Senior high school students, $N = 1,228$; adults, $N = 1,356$; 10th grade students, $N = 451$. Det. = Detachment; Mon. = Money; L.S. = Life Style; Cre. = Creativity; S.D. = Self-Development; Ind. = Independence; Alt. = Altruism; CoW. = CoWorkers.

Measure and procedure. Measure and procedure were the same as in Study 1, except that for the MANOVA, occupations for which subjects were studying or in which they were actually working, were classified according to 22 categories: secretaries, computer programmers, policemen, carpentry apprentices, electrical trades apprentices, motor mechanic apprentices, fitter/machinist apprentices, pastry-cook apprentices, clerks, arts students, science students, engineering students, psychology students, nurses, teaching students, career advisors, psychologists, housewives, managers, personnel officers, hairdresser apprentices, and administrative studies students.

Table 4
Principal Components Analysis With Varimax Rotation for Adults Studying for or Working in Various Occupations

Work aspect preference	Rotated factor loadings		
	Human/personal concern	Non-work orientation	Freedom
Subscale			
Independence	-.21	.35	.71
CoWorkers	.68	.30	.08
Self-Development	.72	-.09	.12
Creativity	.18	-.18	.78
Money	.09	.66	.28
Life Style	.30	.62	.00
Prestige	.46	.33	.32
Altruism	.55	-.12	.21
Security	.63	.46	-.10
Management	.36	.06	.53
Detachment	-.07	.74	-.10
Physical Activity	.63	.30	-.02
Surroundings	.52	.52	.12
Eigenvalue	3.93	1.55	1.38
Percentage of common variance explained	57.20	22.90	19.90
Percentage of total variance explained	30.20	12.00	10.60

Results and Discussion

Factor pattern. Table 4 presents the results of the principal components analysis of the WAPS subscale scores. Three factors with eigenvalues greater than 1.0 were extracted and rotated, accounting for 52.8% of the total variance. As in Study 1, the amount of unexplained variance is high. The first factor, Human/Personal Concern, is characterized by high positive loadings on Self-Development, CoWorkers, Security, Physical Activity, and Altruism. This is a diverse combination of subscales, not easy to relate to previous research findings. However, it does correspond closely with the human/personal-concern cluster found by Pryor (1982). It also seems to be similar to the "non-competitive" grouping suggested by Lofquist and Dawis (1978). Following Pryor (1982, 1983a), this factor was designated Human/Personal Concern.

The second factor, Non-Work Orientation, is characterized by high positive loadings on Detachment, Money, and Life Style. The same combination of loadings found for 11th and 12th grade high school students (see Study 1) was given to these findings for adults.

The third factor, Freedom, has high positive loadings on both Creativity and Independence. These findings appear to be very similar to the Freedom factor found for 11th and 12th grade high school students (see Study 1).

Scoring the factors. The results of the stepwise multiple regression analysis are also shown in the middle part of Table 3. The Rs are all above .90, indicating a satisfactory correspondence between the scores derived from the equations and the factor scores.

Concurrent validity. Using scores derived from the regression equations in Table 3, a MANOVA was performed using the three second-order dimension scores as the dependent variables and occupation category as the independent variable. The significant overall $F(63, 3702) = 10.67, p < .01$, indicated substantial variation in test scores across occupations. Subsequent univariate tests confirmed significant differences in scores for each second-order preference dimension: human/personal concern,

Table 5
Principal Components Analysis with Varimax Rotation for Grade 10 High School Students' Work Aspect Preference Scale Scores

Work aspect preference	Rotated factor loadings		
	Non-work orientation	Power	Human/personal concern
Subscales			
Independence	.33	.56	.16
CoWorkers	.61	.01	.43
Self-Development	.03	.26	.75
Creativity	-.09	.66	.53
Money	.32	.72	-.16
Life Style	.75	.19	.02
Prestige	.22	.74	.28
Altruism	.17	.17	.64
Security	.58	.14	.24
Management	.08	.73	.28
Detachment	.73	.23	-.17
Physical Activity	.41	.01	.50
Surroundings	.65	.24	.31
Eigenvalue	4.60	1.58	1.19
Percentage of common variance explained	62.40	21.30	16.20
Percentage of total variance explained	35.40	12.10	9.20

$F(21, 1242) = 8.45, p < 0.1$; non-work orientation, $F(21, 1242) = 8.95, p < .01$; and freedom, $F(21, 1242) = 10.39, p < .01$. These data support the view that the three dimensions are likely to be valid and potentially useful.

Study 3

Method

Subjects. Subjects were 451 Grade 10 high school students from Victoria, Australia; 55% were boys, and 45% were girls.

Measure and procedure. Measure and procedure were the same as in Study 1, with two exceptions. First, because subjects were not asked their occupational choice, no MANOVA was performed. Second, Ahmavarra's (1957) transformation analysis was used to compare factor patterns found in this study with those of the two previous studies. Reynolds and Harding (1983) did not find substantial differences in the results obtained for six alternative methods of factor comparison. Transformation analysis was used because it takes account of the whole matrix pattern and because it allows rotation to the most congruent solution (Rummel, 1970). In exploratory research, these attributes were viewed as highly desirable. Transformation analysis, however, poses a problem in that its analysis is asymmetrical and therefore requires the specification of a comparison criterion. The criterion matrix nominated in all comparisons was that for the older subjects. This procedure was adopted because research evidence from vocational interests (Strong, 1951) and from the WAPS itself (Pryor & Davies, 1986) indicate that such affective characteristics increase in stability with aging. When the elements of the obtained transformation analysis matrix are normalized, they can be interpreted as the loadings of the younger subjects on the factors of the older subjects (Ahmavarra, 1954).

Results and Discussion

Factor pattern. Table 5 presents the results of the principal components analysis of the WAPS subscale scores for the adult

sample. These factors with eigenvalues greater than 1.0 were extracted and rotated, accounting for 56.7% of the total variance. Again, this suggests that the factorial derivation of the WAPS may have resulted in a large number of very specific factors in the unexplained variance. The first factor shows high positive loadings on Life Style, Detachment, Surroundings, Co-Workers, and Security. In line with previous findings, this factor was designated Non-Work Orientation; it shows similarities with a factor found by Boulton (1980) linking lifestyle and work environment that he characterized as Comfortable Existence.

Highest loadings for the second factor, Power, are for Prestige, Management, Money, Creativity, and Independence. This configuration of loadings suggests a concern for the benefits and prerogatives associated with the exercise of authority. For this reason, although it shows similarities with Freedom in the earlier research, the factor was labeled *Power*. Other research that has found dimensions similar to the Power factor of this study include "status/power" (Pryor, 1980), "competitive"-need category (Lofquist & Dawis, 1978), "power and privilege" cluster (Pryor, 1982), and the management research of England and Lee (1974) and Porter (1961).

The third factor, Human/Personal Concern, has highest loadings on Self-Development and Altruism. It seems to have some similarity with the cluster analysis results of Pryor (1982, 1983a). Factor 3 was tentatively labeled Human/Personal Concern because of its partial resemblance to the Human/Personal Concern factor found for adults (see Study 2).

Scoring the factors. The results of the stepwise multiple regression analysis are set out as equations in the last part of Table 3. All three Rs are .90 or above.

Comparison of factor patterns across samples. The results comparing the factor patterns in Tables 2, 4, and 5, using Ahmavarra's (1954, 1957) transformation analysis, are presented in Table 6. The elements of the transformation matrices can be interpreted as regression coefficients indicating the best (least squares rotation fit) prediction of the column factors in terms of the row factors. When normalized for rows, the elements can be viewed as the loadings of the column factors (younger group) on the row factors (older group). From Table 6, it can be seen that the normalization recommended by Ahmavarra (1957) provides the clearer of the sets of matrix solutions. Subsequent discussion is based on the normalized results.

Inspection of the highest coefficients for each factor in each sample suggests that across the three groups, the same or similar three factors appear to be present for each group. Furthermore, the pattern of the elements of each matrix appears to provide a reasonable approximation to simple structure. Thus, each of the Grade 10 student factors shows one loading above .85 with one adult factor and no longer loadings above .40 on other adult factors, a matrix pattern duplicated in the Grade 11/12 Students \times Adults matrix. The matrix for Grade 10 Students \times Grades 11/12 Students has two secondary loadings greater than .40, but overall it approximates simple structure.

The results from Table 6 provide strong support for the similarity of the patterns of factors across the three groups. This suggests that the structure of general work preference dimensions appears to be fairly generalized. It can further be inferred from these results that each of the general work preference dimensions is recognizable in each of the matrices, which provides further support for their stability across samples. For

Table 6
Transformation Analyses Comparing Patterns of Factor Loadings Across Three Samples

Factor		Transformation matrices					
Grade 10 students							
Adults	NWO	POW	H/PC	NWO	POW	H/PC	
H/PC	.29	.08	.77	.35	.10	.93	
NWO	.87	.37	-.31	.87	.37	-.31	
Fre	-.05	.97	.38	-.05	.93	.36	
Grade 11/12 students							
Adults	NWO	FRE	PC	NWO	FRE	PC	
H/PC	-.05	-.05	.81	-.06	-.06	1.00	
NWO	1.05	.24	.05	.97	.22	.05	
Fre	.26	.92	.04	.27	.96	.04	
Grade 10 students							
Grade 11/12 students	NWO	POW	H/PC	NWO	POW	H/PC	
NWO	.78	.37	-.26	.87	.41	-.29	
Fre	-.14	.53	.17	-.24	.92	.30	
PC	.42	.11	.85	.44	.12	.89	

Note. Columns 1, 2, and 3 of the transformation matrices are unnormalized; columns 4, 5, and 6 are normalized. NWO = Non-Work Orientation; POW = Power; H/PC = Human/Personal Concern; FRE = Freedom; PC = People Concern.

Non-Work Orientation, this inference appears clearly justified in the similarity of subscale predictors across groups (see Tables 2, 4, and 5). For Human/Personal Concern, the inference is substantially confirmed for Grade 10 students and adults, but less so for Grades 11/12 students, having only Altruism in common with the other two regression approximations (see Tables 2, 4, and 5). For Freedom, the case is similar to Human/Personal Concern in that the inference of stability appears supported for Grades 11/12 students and adults, but less so for Grade 10 students, given the differences in predictors in the multiple regression approximations. Considering all of these results, it can be concluded that the structure of general work preference dimensions is clearly similar across samples but that in this structure some differences in dimensional content and character are discernible, which justify specific differences in dimension designation in some instances (e.g., power and personal concern).

Study 4

Method

Subjects. Data used were from two samples of senior high school students—Grades 11 and 12. There were equal numbers of boys and girls in each group; 130 in Sample 1, and 140 in Sample 2.

Measure and procedure. The WAPS was administered twice to Sample 1, with a 2-week period between each testing. Similarly, the WAPS was administered to Sample 2, with a 6-week period between testings. For both groups, the WAPS data were scored according to equations in Table 3, to derive second-order preference dimension scores. For each sample, these results were correlated. For the same dimension, the correlation of the two scores constitutes a test-retest reliability coefficient.

Results and Discussion of Study 4

Short-term reliability. The results for Sample 1 were as follows: Non-Work Orientation, $r_{tt} = .77$; Freedom, $r_{tt} = .84$; and Human/Personal Concern, $r_{tt} = .78$.

Medium-term reliability. The results for Sample 2 were as follows: Non-Work Orientation, $r_{tt} = .75$; Freedom, $r_{tt} = .78$; Human/Personal Concern, $r_{tt} = .80$.

Psychometricians widely agree that a reliability coefficient of .80 or above is desirable for cognitive measures. It is, however, fairly generally conceded that noncognitive measures cannot realistically be expected always to meet this criterion (Cronbach, 1970; Kline, 1979). An adequate reliability coefficient for such tests is usually set at about .70.

In light of these conventions and the fact that test-retest reliability constitutes the most stringent of methods to assess reliability (Tuckman, 1975), it can be concluded that the correlations obtained in both studies of the reliability of the WAPS general preference dimensions are close to or exceed the cognitive test criterion.

Although it was expected that the medium-term reliability coefficients were likely to be lower due to individuals' changing over the longer period, the results indicate only a small attenuation of the correlations. Both sets of results suggest that the WAPS general work preference dimensions are sufficiently reliable to be used for both research and counseling purposes.

General Discussion

To account adequately for individual differences, we need to investigate how people may differ across dimensions as well as along them. In the field of values and preferences related to work, such investigation has been limited by a variety of shortcomings in research design and by the general failure actually to assess more general dimensions and use them as a basis for subsequent research. The current study is intended as a first step to redress this situation. It is based on the premise that simply proliferating multivariate studies of inventories of values and preferences related to work is insufficient. Rather, dimensions have to be measurable at different levels of generality.

This research endeavors to do this. It has been demonstrated that in addition to the 13 multivariately derived first-order factors of the Work Aspect Preference Scale, three similar sets of second-order factors across varying samples are identifiable. This generalizability of underlying preference dimensions is based on the similarity of factor patterns across the three samples in this study. However, this similarity is not complete. Further research is required to indicate whether particular divergences in general work preferences for the current three groups are substantive or attributable to chance variation. Such further research may also reveal whether a single set of factor pattern loadings may be sufficient to represent the general work preferences for these three groups.

From the present data, scoring procedures for each general work preference dimension derived were developed using stepwise multiple regression to provide an accurate and practicable way to use the results of this research. The final two steps of this research relate to the central measurement issues of validity and reliability. The concurrent validity of the three dimensions was supported by the findings for two distinct samples. The scores of all three dimensions significantly varied across either actual occupations or occupational choice categories. Furthermore, these data (Pryor, 1986) indicate generally expected patterns of high and low scores on the dimensions. For example, those highest on human/personal concern are nurses and aspirants to social and paramedical sciences; farmers and computer programmers are lowest. For non-work orientation, clerks are highest, whereas psychologists are lowest. For freedom, pastry-cook apprentices and creative arts aspirants are highest; policemen and nurses are lowest. In terms of the reliability of the three general work preference dimensions, two test-retest reliability studies were conducted that suggested that for both short- and medium-term reliability, these dimensions are adequate for both research and counseling purposes.

The three general work preference dimensions may be useful in several ways. First, because the WAPS was developed principally for use in the context of vocational counseling and career development, the application in this area is obvious. The concurrent validity data support this use of the dimensions. Second, the dimensions may be useful in some situations including staff selection and allocation of duties. For example, one employer is currently experimenting with the human/personal concern dimension to assist in the selection of staff for duties dealing with people because the employer has had serious public relations problems with clerical staff who were not people oriented.

Third, Doring (1984), after a review of relevant research studies, concluded that the experience of unemployment is much more devastating for those who are highly committed to work. Therefore, it can be expected that low scorers on the non-work orientation dimension are likely to be in this at-risk category. Such information would be useful for understanding the dynamics of the unemployment experience and may suggest ways that counselors may help clients cope with it (Pryor & Ward, 1985). Fourth, because, as Youngblood (1984) has indicated, attachment to work is inversely related to absenteeism, the non-work orientation dimension, if so validated, could be used in the selection of staff to reduce this employer cost.

Further research could focus on the derivation and measurement of more general value/preference dimensions, using in-

ventories other than the WAPS. There also appears to be a need to pursue the measurement of general (i.e., global as distinct from specific) and generalizable (i.e., widespread rather than sample particular) patterns of subscales derived from multivariate techniques other than factor analysis, like cluster analysis. With respect to the dimensions found in this research, more work needs to be done to establish their construct validity. For example, what is the difference, if any, between the human/personal concern dimension and the social service scales of many interest inventories? Obviously, more research is required to provide further evidence about the value of assessing general work preference dimensions. However, even if in the long term such dimensions are not of much value, the information would still be an improvement over the piecemeal conglomeration of research findings that currently constitute the study of different levels of generality of values and preferences related to work.

References

- Ahmavarra, Y. (1954). Transformation analysis of factorial data. *Annales Acadamie Scientiarum Fennicae*, 88, 1-150.
- Ahmavarra, Y. (1957). On the unified factor theory of mind. *Annales Acadamie Scientiarum Fennicae*, 106, 1-176.
- Athanasou, J. A. (1986). Factor analysis of the vocational interest scales of the Holland vocational interest scales of the Holland Vocational Preference Inventory. In Taylor, K. F. & Lokan, J. J. (Eds.), *Holland in Australia: A vocational choice theory in research and practice*. Hawthorne (Victoria): Australian Council for Educational Research.
- Beynon, H., & Blackburn, R. M. (1972). *Perceptions of work: Variations within a factory* (Cambridge Papers in Sociology No. 3). London: Cambridge University Press.
- Boulton, B. (1980). Second-order dimensions of the Work Values Inventory (WVI). *Journal of Vocational Behavior*, 17, 33-40.
- Campbell, D. P., & Hansen, J. C. (1981). *Manual for the Strong Campbell Interest Inventory* (3rd ed.). Stanford, CA: Stanford University Press.
- Cattell, R. B. (1946). *The description and measurement of personality*. New York: World Book.
- Cattell, R. B. (1973). *Personality and mood by questionnaire*. San Francisco: Jossey-Bass.
- Cook, J. D., Hepworth, S. J., Wall, T. D., & Warr, P. B. (1981). *The experience of work*. London: Academic Press.
- Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.
- Doring, A. J. (1984). Beliefs about the employed and the unemployed held by senior high school students in North Queensland. *Australian Journal of Education*, 28, 78-88.
- Dubin, R., & Champoux, J. E. (1975). Worker's central life interests and personality characteristics. *Journal of Vocational Behavior*, 6, 165-174.
- England, A. W., & Lee, R. (1974). The relationship between managerial values and managerial success in the United States, India, and Australia. *Journal of Applied Psychology*, 59, 411-419.
- Eysenck, H. J., & Eysenck, S. B. G. (1969). *Personality structure and measurement*. San Diego, CA: Knapp.
- Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire*. Kent, England: Hodder & Stoughton.
- Goldthorpe, J. H., Lockwood, D., Beckhofer, P., & Platt, J. (1968). *The affluent worker: Industrial attitudes and behavior*. Cambridge, England: Cambridge University Press.
- Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.
- Guilford, J. P. (1959). *Personality*. New York: McGraw-Hill.
- Guilford, J. P. (1975). Factors and factors of personality. *Psychological Bulletin*, 82, 802-814.

- Hendrix, V. L., & Super, D. E. (1968). Factor dimensions and reliability of the work values inventory. *Vocational Guidance Quarterly*, 17, 269-274.
- Holland, J. L. (1973). *Making vocational choices: A theory of careers*. Englewood Cliffs, NJ: Prentice-Hall.
- Kline, P. (1979). *Psychometrics and psychology*. New York: Academic Press.
- Lofquist, L. H., & Dawis, R. V. (1978). Values as second-order needs in the theory of work adjustment. *Journal of Vocational Behavior*, 12, 12-19.
- Mortimer, J. (1975). Occupational value socialization in business and professional families. *Sociology of Work and Occupations*, 2, 29-53.
- O'Connor, J. P., & Kinnane, J. F. (1961). A factor analysis of work values. *Journal of Counseling Psychology*, 8, 263-267.
- Porter, L. W. (1961). A study of perceived need satisfactions in bottom and middle management jobs. *Journal of Applied Psychology*, 45, 1-10.
- Prediger, D. J. (1976). A world of work map for career exploration. *Vocational Guidance Quarterly*, 24, 198-208.
- Pryor, R. G. L. (1979). In search of a concept: Work values. *Vocational Guidance Quarterly*, 27, 250-258.
- Pryor, R. G. L. (1980). Some types of stability in the study of students' work values. *Journal of Vocational Behavior*, 16, 146-157.
- Pryor, R. G. L. (1981a). Tracing the development of the Work Aspect Preference Scale. *Australian Psychologist*, 16, 241-257.
- Pryor, R. G. L. (1981b). Interests and values as preferences. *Australian Psychologist*, 16, 258-272.
- Pryor, R. G. L. (1982). Values, preferences, needs, work ethics and orientations to work: Toward a conceptual and empirical integration. *Journal of Vocational Behavior*, 20, 40-52.
- Pryor, R. G. L. (1983a). Sex differences in the levels of generality of values/preferences related to work. *Journal of Vocational Behavior*, 23, 233-241.
- Pryor, R. G. L. (1983b). *Manual for the Work Aspect Preference Scale*. Melbourne: Australian Council for Education Research.
- Pryor, R. G. L. (1986). *Toward the measurement of General Work Preference Dimensions* (Research Rep.). Darlinghurst, New South Wales, Australia: N.S.W. Department of Industrial Relations.
- Pryor, R. G. L., & Davies, R. (1986). Two studies of the reliability of the Work Aspect Preference Scale for tertiary samples. *ACER Bulletin for Psychologists*, 40, 8-10.
- Pryor, R. G. L., & Ward, R. T. (1985). Unemployment: What counselors can do about it. *Journal of Employment Counseling*, 22, 10-31.
- Reynolds, C. R., & Harding, R. E. (1983). Outcome in two large sample studies of factorial similarity under six methods of comparison. *Educational and Psychological Measurement*, 43, 723-728.
- Rummel, R. J. (1970). *Applied factor analysis*. Evanston, IL: Northwestern University Press.
- Strong, E. K. (1951). Permanence of interest scores over 22 years. *Journal of Applied Psychology*, 36, 65-74.
- Taylor, N. B., & Pryor, R. G. L. (1986). The conceptualization and measurement of preferences: Vocational and work aspect. *British Journal of Guidance and Counselling*, 14, 66-77.
- Tuckman, B. W. (1975). *Measuring educational outcomes*. New York: Harcourt-Brace Jovanovich.
- Vernon, P. E. (1979). *Intelligence: Heredity and environment*. San Francisco: Freeman.
- Youngblood, S. A. (1984). Work, non-work, and withdrawal. *Journal of Applied Psychology*, 69, 106-117.

Received May 22, 1986

Revision received December 8, 1986

Accepted December 15, 1986 ■

Effects of Using High- Versus Low-Performing Job Incumbents as Sources of Job-Analysis Information

Patrick R. Conley
Chicago Police Department

Paul R. Sackett
University of Illinois at Chicago

We investigated the relation between incumbent performance level and job-analysis information using three different methods. Separate groups of high- and low-performing incumbents generated lists of tasks and of knowledges, skills, and abilities (KSAs) for the job of youth officer for a large metropolitan police force. These lists were virtually the same for all groups. In addition, those tasks and KSAs omitted by any of the groups were found to be unimportant in later ratings. Group differences in the ratings of the inventories on a number of scales as well as the ratings of the 18 cognitive Fleishman Ability Scales by the entire incumbent population ($N = 179$) and their supervisors ($N = 9$) were examined. The individual scales were factor analyzed, and discriminant analyses were applied to the factor scores to identify any differences in the ratings of high and low performers and supervisors. No differences were found. The limitations of this study are discussed.

Recently, procedural issues associated with job analysis have been investigated in several studies. Issues include the effects of different job-analysis methods (Cornelius, Carron, & Collins, 1979; Levine, Ash, & Bennett, 1980) and the quality of ratings made by "job knowledge experts" (Ash, Levine, Higbee, & Sis-trunk, 1983). This study expands these earlier studies to include the effect of using high- and low-performing incumbents as the source of job-analysis information.

The relation between incumbent performance levels and job-analysis data can be hypothesized to be influenced by the individual's perception of the job. In jobs that allow the workers the discretion to place emphasis on any of a number of job activities, one might expect individuals to have different perceptions of the job's demands. High and low performers may differ in their choice of the time spent in various activities and in their insight into the knowledges, skills, and abilities (KSAs) that result in high performance.

In the one study we found that focused on the relation between performance level and job-analysis information, Wexley and Silverman (1978) investigated responses of retail store managers to a structured job-analysis questionnaire. The questionnaire was developed for managerial positions and focused on 7 work activities (e.g., personnel and credit management, customer and community relations, etc.) and 30 worker characteristics (e.g., knowledge of company, decisiveness, decision making, organizing and planning, etc.). The questionnaire was given to 146 managers, who were asked to rate (a) each of the worker activities on scales of importance and time spent and (b) the worker characteristics on a scale of importance. The store

managers classified as effective or ineffective did not differ significantly in their ratings.

This study goes beyond the Wexley and Silverman study in several ways. First, it focused on a complex nonmanagerial job: a juvenile officer in a police organization. Second, it asked for ratings of task statements on a number of scales as well as ratings for three KSA scales (Primoff, 1975); it also included the Fleishman Scales of cognitive abilities (Theologus, Romasko, & Fleishman, 1970). Finally, it examined the relation between incumbent job-performance level and the generation of task and KSA inventories. To date there has been no systematic investigation into the effects of performance-level differences on the generation of either task or KSA inventories.

Method

Setting

This study was part of a larger program designed to develop a selection instrument for the position of juvenile officer for the Chicago Police Department. Juvenile officers are responsible for investigations in which there are either young offenders (i.e., less than 17 years of age) or youthful victims (e.g., child-abuse cases, child pornography, etc.).

Subjects

Subjects in this study consisted of all active-duty juvenile officers ($N = 179$) and their supervisors ($N = 9$) assigned to actual field investigations within the Youth Division of the Chicago Police Department. Table 1 describes the number of subjects in each of the various groups. Although the manner in which subjects were classified into groups will be discussed below, several comments are needed concerning those subjects classified under the "other" category.

Initially all officers and supervisors were scheduled to participate in all rating sessions. Unfortunately, a number of officers completed one or more of the scales but did not complete all ratings. During the 7-month data-collection phase, a number of officers were routinely given assignments that required absence from their units (i.e., undercover and out-of-town investigations). In addition, a number of officers were transferred from the division.

An earlier version of this article was presented at the 93rd Annual Convention of the American Psychological Association, August 1985, Los Angeles, California.

Correspondence concerning this article should be addressed to Patrick R. Conley, Chicago Police Department, 1300 W. Jackson, Chicago, Illinois 60605.

Table 1
Numbers of Officers Completing Ratings of Job-Analysis Data

Group	N	%
High performers	38	33.3
Low performers	34	29.8
Medium performers	33	28.9
Supervisors	9	7.9
Subtotal	114	100.0
Other	74	
Total number of officers	188	

Note. All *N*s represent officers who completed all ratings, except for *Other*, which represents officers who provided partial data. Total percentage is rounded.

Of the 188 officers who participated, 114 completed the full sequence of ratings. Although all ratings were used in the factor analyses of rating data to ensure the stability of factor loadings, only data from raters who completed all ratings were used in the group comparisons.

Procedure

Group identification. Incumbent performance groups were derived by using a departmental performance-rating system that requires the biannual ranking of individual officers. The ranking is determined in meetings (in each of the city's six police areas) in which field supervisors discuss monthly attendance and production data (e.g., medical absences, arrests, cases handled, etc.) and rank order their subordinates accordingly. Thus, there are six different employees that receive the top ranking. This ranking becomes the officer's performance rating for the 6-month period.

Performance ratings of all current youth officers assigned to field investigation were obtained for the last five rating periods, and the average of the highest four was computed. This averaging process has been used by the department and is based on the philosophy that one deviant rating should not exclude a person from promotion. High- and low-performance groups were defined as those officers in the highest and lowest thirds of the distribution of average performance grades.

In order to estimate the system's reliability, intercorrelations of the last five efficiency ratings were computed. These ratings, which spanned a 30-month period, yielded an average intercorrelation of .73, with a range from .57 to .88, indicating a reasonably reliable rating system. A correlation of .91 was found between two consecutive ratings of 33 officers transferred between units within the division, suggesting that high reliability is not simply the result of memory effects on the part of raters.

We attempted to address two concerns about this rating procedure. The first was the possibility that differences between the six areas are so great that individuals ranked as top performers in one area would be low performers in another and vice versa. The second was that the ratings may not represent differences between incumbents in job-relevant behavior, but may merely reflect the biases and personal preferences of the supervisors providing the ratings. Although objective performance measures would have been desirable either in lieu of or in addition to supervisory ratings, records of indices such as the number of investigations handled and the number of arrests made are not routinely kept and thus were not available. We were able to obtain measures of investigations completed per month and arrests made per month for a 3-month period. In order to control for sick days and vacation, both measures were adjusted for number of days worked per month. Table 2 reports

these indices for the six highest- and six lowest-ranked officers in each of the six areas. The results clearly indicate that in each area high performers complete both more investigations and are responsible for more arrests than their low-performing peers. Low performers achieved essentially the same level of objective performance regardless of area. Although there is variability between areas in the objective performance of high performers, differences between high and low performers are substantial in all areas. Thus, there is evidence that the supervisor ratings reflect real behavioral differences and that the findings are not confounded by lack of comparability between areas.

Task generation and rating. In order to compare the quality of the inventories generated from the different performance groups, two groups of high performers and two groups of low performers, each group consisting of four incumbents, generated preliminary inventories. Each group meeting lasted 1 to 1½ days.

The four groups generated separate inventories that contained 451 tasks in total. The senior author and an assistant personnel officer unaware of the purpose of the study each independently eliminated redundant tasks and collated the separate inventories into a composite inventory. These two independently derived inventories had 108 tasks in common; the senior author's had 3 additional tasks. A final master inventory consisting of all nonredundant tasks ($N = 111$) was then constructed. In order to compare relative quality of the inventories generated by each high- and low-performing group, the proportion of tasks overlapping the master inventory was computed for each group.

The composite master task inventory was then administered to all available officers within the division. This included all field investigators and supervisors (including those officers who generated the various preliminary inventories). In total, 154 of the 188 officers participated (i.e., 145 incumbents and 9 supervisors). Subjects rated tasks on each of four 7-point scales including: (a) importance, (b) relative difficulty, (c) relative time spent, and (d) criticality of error. Officers were instructed to complete ratings for each of the 111 tasks on a single scale before proceeding to the next.

KSA generation and rating. The next step involved the generation of KSAs needed on the job. This was done by using two groups of high performers and two groups of low performers (using different individuals from those used in the task-generation phase, and again using 4 individuals per group) to develop the KSAs needed to perform the tasks rated as important. (Tasks were defined as important if 60% of incumbents rated the task as a 5 or above on either the importance or the criticality scale.) Each meeting took one full day. Using the same method as was used with the task data, a composite list of 32 KSAs was compiled. The separate lists generated by each group were then compared with the master KSA list to determine the amount of overlap. The KSA master list was then rated by all available field personnel ($N = 138$). Ratings were made on three scales that were collapsed to an "item index" representing the relative importance of an ability. This index is represented by the formula $I = S \times P + T$, in which I is the item index, S is a rating of the ability's potential to identify superior workers, T is the rating of the ability's necessity on the job, and P is an estimate of the number of candidates having this ability. Primoff (1975) provides rationale for this formulation.

Fleishman Scales rating. The 18 cognitive scales of the Fleishman Ability Scales (Theologus et al., 1970) were used. A prestudy investigation determined that the rating of the 10 additional scales of manual abilities added little variance to the individual profiles.

Data analysis. A principal-components analysis was performed on each scale, using the data from all subjects rating the particular scale. Although the preferred method would have been to compare the factor structure of each group, the limited ratio of subjects to tasks prohibited such analysis. Next, a scree test (Cattell, 1966) determined the number of factors, after which a varimax rotation helped identify interpretable

Table 2
Mean Performance Levels for Each of Two Objective Performance Indices for High and Low Performers Assigned to the Six Police Areas

Performance group/measure	Police area						Total
	1	2	3	4	5	6	
High							
Cases processed							
<i>M</i>	36.40	34.33	50.18	49.73	45.74	39.83	42.70
<i>SD</i>	6.07	5.16	12.18	9.63	7.94	4.67	9.86
Arrest index							
<i>M</i>	7.71	19.86	6.94	7.29	16.16	16.48	12.40
<i>SD</i>	3.29	9.94	3.51	2.38	8.16	4.21	7.64
Performance composite							
<i>M</i>	44.10	54.20	57.12	57.03	61.90	56.31	55.11
<i>SD</i>	8.33	10.06	13.46	11.51	14.23	7.16	11.66
Low							
Cases processed							
<i>M</i>	22.92	19.08	20.44	20.98	26.76	23.69	22.31
<i>SD</i>	3.81	4.99	7.58	4.84	4.48	3.19	5.29
Arrest index							
<i>M</i>	4.53	12.83	2.23	3.42	6.89	6.05	5.99
<i>SD</i>	1.71	6.69	1.44	1.25	2.84	5.25	4.95
Performance composite							
<i>M</i>	27.46	31.91	22.67	24.41	33.66	29.75	28.31
<i>SD</i>	4.39	6.90	7.35	6.00	3.72	4.71	6.57

factors. Factor scores were computed, after which discriminant analyses were conducted to determine group differences.

Results

The Quality of Task and KSA Inventories

When the group-generated task and KSA inventories were compared to the master inventories, no differences were found. For the two high-performance groups, 97% and 99% of the tasks overlapped the master inventory; for the low-performance groups, the figures were 95% and 97%. Similarly, the high performers' KSA inventories agreed with the master KSA list at rates of 94% and 97%; whereas low-performance groups matched at 91% and 94%.

In addition, none of the tasks or KSAs omitted by any of the generation groups were rated as either important or critical in the subsequent rating phase.

Group Ratings

The number of interpretable factors and the amount of variance accounted for varied according to the scale. For example, six-factor solutions were obtained for the importance scale (i.e., juvenile procedures, adult arrests, missing investigations, decision making, interviewing, and miscellaneous duties; accounting for 51% of the total variance), time spent scale (i.e., juvenile procedures, theft investigations, decision making, missing investigations, interviewing and reporting procedures; representing 53.5% of total variance), and the KSA index (i.e., knowledge of department procedures, on-the-job skills, communication skills, specialized knowledge, socioeconomic knowledge, and interviewing; 56.2%). Interpretable five-factor solutions were obtained for both the criticality and difficulty scales (account-

ing for 54.6% and 57.8% of total variance), and a four-factor solution was identified in the Fleishman Scales (representing 63.4% of the total variance). Space consideration precludes detailed discussion of these results; details of the principal-components analyses are available from the senior author.

When differences between high and low performers and supervisors were assessed by use of the discriminant-analysis procedure, no differences were found. Table 3 lists the results of these analyses for the four task scales as well as the KSA item index and the Fleishman Ability Scales.

Discussion

This study investigated the relation between incumbent performance levels and job-analysis information for each of three

Table 3
Tests of Centroid Differences for Each Scale of the Task Scales, KSA Item Index, and Fleishman Scales

Scale	Wilks's λ	Approximate <i>F</i>
Importance	.874	.95 <i>ns</i>
<i>df</i>	(6, 2, 78)	(12, 148)
Time Spent	.754	1.02 <i>ns</i>
<i>df</i>	(6, 2, 78)	(12, 148)
Criticality	.777	1.90 <i>ns</i>
<i>df</i>	(5, 2, 78)	(10, 148)
Difficulty	.812	1.61 <i>ns</i>
<i>df</i>	(5, 2, 78)	(10, 148)
KSA item index	.824	1.50 <i>ns</i>
<i>df</i>	(5, 2, 78)	(10, 148)
Fleishman Abilities	.937	.61 <i>ns</i>
<i>df</i>	(4, 2, 74)	(8, 150)

different job-analysis methods. Groups of high- and low-performing incumbents from a single job (i.e., juvenile officers from a large metropolitan police department) generated task and KSA inventories that were subsequently rated by the entire incumbent and supervisor population. The 18 Fleishman Ability scales were also rated.

The results also indicated that there were no differences between the incumbent groups in terms of the generation of either task information or KSAs needed in the performance of the job. Comparisons of individual group lists with a composite inventory indicated that both groups were able to identify both types of information. In addition, the tasks and KSAs omitted by one or the other groups during the generation phase were later rated as unimportant.

Nor were group differences found when the ratings of the various groups were examined. High performers, low performers, and supervisors rated each of the various scales of the task inventory, the KSA item index and the Fleishman Ability Scales in a similar manner. Discriminant analysis from each of the scales' factor scores indicated that there were no differences in the ratings from the various groups.

In examining the results of this study, the possibility of a Type II error should be considered. Group differences may have been obscured by the consolidation of the groups' ratings into a single data base from which factor scores were computed. The size of the incumbent population prohibited factoring high- and low-performers' ratings separately. Thus, real group differences may have been obscured. In addition, true differences in the task ratings may have been masked by the stability of the factor structure resulting from the limited number of incumbents relative to the total number of tasks rated. Although we recognize that the ratio of subjects to variables is low, what must be remembered is that this study relied on ratings from the entire incumbent population. Thus, the results represent the maximum amount of information available.

The lack of findings may also reflect the impact of environmental influences on the ratings. For example, the amount of preservice training received by juvenile officers may have mediated the perception of job requirements. In this case, newly selected officers are required to attend a month-long, preservice school that describes the job in terms of responsibilities and

duties, provides instruction in specialized topics (e.g., juvenile law, reporting procedures, etc.), and details performance requirements. Thus, the finding of no differences in the quality of job-analysis information may be restricted to those organizations that provide a more structured view of job requirements.

In all, the implications of this study reflect previous research. In agreement with the conclusions of Wexley and Silverman (1978), this study found no differences in the quality of job-analysis data of high or low performers either when (a) the type of method used varied or (b) incumbents actually generated the various types of inventories. Thus, at least in this situation where there has been extensive training, the performance level of the incumbent selected to give job-analysis information appears to make little difference in the results obtained. However, in those situations in which the job under examination has little structure or includes little or no formal training, the practitioner should be aware of the possibility of reaching contradictory conclusions, depending on the incumbents selected to give this information.

References

- Ash, R. A., Levine, E. L., Higbee, R. H., & Sistrunk, F. L. (1983, August). *Comparison of task ratings from subject matter experts versus job incumbents*. Paper presented at the 91st Annual Convention of the American Psychological Association, Anaheim, CA.
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245-276.
- Cornelius, E. T., Carron, T. J., & Collins, M. N. (1979). Job analysis models and job classification. *Personnel Psychology*, 32, 639-708.
- Levine, E. L., Ash, R. A., & Bennett, N. (1980). Exploratory comparative study of four job analysis methods. *Journal of Applied Psychology*, 65, 524-535.
- Primoff, E. S. (1975). *How to prepare and conduct job element examinations*. Washington, DC: U.S. Government Printing Office.
- Theologus, G. C., Romasko, T., & Fleishman, E. A. (1970). *A feasibility study of ability dimensions for classifying human tasks*. Washington, DC: American Institute for Research.
- Wexley, K. N., & Silverman, S. B. (1978). An examination of differences between managerial effectiveness and response patterns on a structured job analysis questionnaire. *Journal of Applied Psychology*, 63, 646-649.

Received July 24, 1986

Accepted November 18, 1986 ■

Method Variance as an Artifact in Self-Reported Affect and Perceptions at Work: Myth or Significant Problem?

Paul E. Spector
University of South Florida

Method variance is an artifact of measurement that biases results when relations are explored among constructs measured by the same method. The existence of method variance was explored for affective and perceptual constructs frequently used in organizational research. Data from multitrait-multimethod analyses, studies of social desirability and acquiescence, and relation of self-report and records of absenteeism were presented. Little evidence for method variance as a biasing problem was found with these measures. I conclude that properly developed instruments of the type studied here are resistant to the method variance problem, but that validity of these instruments cannot be assumed on the basis of these results.

Method variance is a potential problem of frequent concern to researchers in areas of psychology that are heavily dependent on questionnaire methods. Method variance is considered to be an artifact of measurement that biases results when relations are explored among constructs measured in the same way. Campbell and Fiske (1959) described method variance as variance attributable to measurement method rather than to the variables or constructs of interest. Campbell and Fiske noted the following common examples: apparatus factors with laboratory animals, halo effects in ratings, and response sets in self-report questionnaires. There seems to be widespread belief among researchers in the industrial/organizational (I/O) area that method variance accounts for considerable shared variance among self-report measures. The adage, generally stated facetiously, that all self-report measures intercorrelate at .30 reflects this belief. One might expect to find a body of research on this problem, but surprisingly little has been written about it. My purpose is to explore the extent of the problem, using both published data and data that I have collected. The treatment will be limited to self-report measures of affect (primarily job satisfaction) and perceptions of the work environment.

Nature of the Problem

Method variance concerns variance in measurement attributed to the particular instrumentation rather than to the construct of interest. Instrument- or method-specific bias accounts for method variance in that common bias sources will be correlated and may produce spurious results. For example, halo bias is common in performance appraisals. Raters tend to rate each ratee the same across different dimensions of performance. Intercorrelations among dimension scores can be attributed to intercorrelations among a common source of bias rather than to relations among true scores of performance dimensions.

The most frequently found sources on method variance in

self-reports were concerned with acquiescence and social desirability (SD). Acquiescence is the tendency for a respondent to agree with items, regardless of content. Acquiescence can also be negative, with a respondent disagreeing with all of the items, regardless of content. Social desirability is the tendency for a respondent to choose the socially desirable response, regardless of the veracity of that response. Both sets have been frequently researched and have long histories of development in psychological testing.

Acquiescence was studied extensively in the 1940s and 1950s. Cronbach (1950) summarized evidence for the existence of the bias and made suggestions for designing tests to minimize it. Such bias is most likely to occur when test items are ambiguous or when tests are poorly developed (Cronbach, 1950). Interest in acquiescence seems to have diminished in the wake of Husek's (1961) work, which showed that various measures of acquiescence were uncorrelated with one another. His conclusion was that there does not exist a specific acquiescence trait common across instruments. If this conclusion is true, one would not expect acquiescence bias components to correlate across instruments.

Social desirability sets are of particular concern in the measurement of personality and psychopathology. Considerable research has shown that some measures are highly correlated with social desirability, making it unclear whether they actually measure their intended constructs (Nunnally, 1978, p. 557). Developers of self-report instruments often screen out items that correlate with social desirability, and of course, forced-choice item formats were developed to deal with this bias. It would seem that social desirability is potentially an important source of method variance in measures of self-reports. It is interesting that little research has been conducted with it in the measurement of organizational variables.

Although many researchers assume that method variance is a cause of spuriously high correlations between variables measured by self-report, biases of the type discussed here may have other effects as well. James, Demaree, and Wolf (1984) pointed out how bias can disrupt score distributions, making them non-normal in shape. Biases can also have attenuating or moderating effects (Ganster, Hennessey, & Luthans, 1983). As Ganster

Correspondence concerning this article should be addressed to Paul E. Spector, Department of Psychology, University of South Florida, Tampa, Florida 33620.

et al. pointed out with social desirability, if SD correlates with both variables of interest, it will artifactually inflate their relationship. If SD correlates with only one of two variables, it will act as a suppressor and artifactually reduce the correlation between them. Finally, individuals high in SD may exhibit a relationship between two variables that is different from that of individuals low in SD.

There is one caution in interpreting social desirability as a response bias. Considerable research has investigated social desirability as a personality construct in its own right (Crowne & Marlowe, 1964). Correlations of variables with social desirability may well reflect relations between two constructs rather than between a construct and source of bias. For example, if job satisfaction were shown to correlate with social desirability it might be because the satisfaction items were subject to SD bias. Alternatively, the correlation might be observed because individuals high in SD tend to be satisfied with their jobs. Merely demonstrating a correlation with SD does not mean automatically that bias has occurred. Additional validation and research would be necessary to draw such a conclusion. For example, Arnold, Feldman, and Purbhoo (1985) found that SD moderates the relation between turnover and affective variables. Whether this result can be attributed to bias or to a personality construct is open to question.

Overlap of item content is a methodological problem that many researchers consider a form of method variance (e.g., Aldag, Barr, & Brief, 1981; Roberts & Glick, 1981). Overlap occurs when two instruments contain identical or very similar items. Job satisfaction instruments, for example, usually contain evaluative as well as descriptive items. Often these descriptive items will be similar to items in measures of job characteristics. The problem can arise from carelessness in the development of instruments, suggesting poor validity, or from overlap in the constructs themselves. An example is the overlap between organizational commitment and intention of quitting the job. One of the major components of commitment is desire to remain on the job. Instruments designed to measure commitment (e.g., Mowday, Steers, & Porter, 1979) include items measuring this aspect. It is not surprising that organizational commitment correlates with intention of quitting the job (e.g., $-.60$ in Michaels & Spector, 1982). This result should not be considered a case of method variance, as discussed here, but rather as an example of construct overlap. (For further discussion of this problem, see Nicholls, Licht, & Pearl, 1982.)

Evidence for Method Variance

As was previously discussed, there are few discussions in the I/O literature on method variance, despite the widespread belief about its prevalence. In addressing this problem, I searched my own files as well as the relevant I/O literature. I conducted a manual search using *Psycscan: Applied* (1983–1985) and five I/O journals, *Journal of Applied Psychology* (1973–1985), *Personnel Psychology*, *Academy of Management Journal*, *Journal of Occupational Psychology*, and *Organizational Behavior and Human Performance* (all 1976–1985). Much of this evidence was found in articles concerned with the validity or development of instruments rather than method variance itself. The three major categories of evidence were from multitrait-multimethod matrices, correlations of relevant variables with social

desirability, and correlations of relevant variables with measures of acquiescence.

Only self-report measures of perceptions of jobs and work environments and affective reactions to jobs were investigated. Measures of abilities, personality, and performance ratings were not included because much has been written about their problems and limitations.

Multitrait–Multimethod Matrices

Campbell and Fiske (1959) have provided a procedure for detecting method variance with their multitrait-multimethod analysis. This procedure provides a global test for method variance, but does not allow the detection of the source or type of bias. These analyses can be conducted when at least two constructs are measured by at least two methods. The pattern of intercorrelations among the measures allows for a test of convergent validity (the tendency of alternate measures of the same construct to correlate strongly), discriminant validity (the tendency for measures of different constructs to correlate, at most, moderately), and method variance (the tendency for different traits measured with the same method to correlate more highly than different traits measured with different methods).

Before a test for method variance can be conducted, one must demonstrate convergent validity. If different constructs are being measured by different methods, it is not possible to distinguish method variance from discriminant validity (Marsh, 1983). For example, suppose that two methods, A and B, are used to assess two traits, 1 and 2. Convergent validity will produce strong correlations between the same traits measured by different methods, or the A_1 versus B_1 , and A_2 versus B_2 correlations will be high. Method variance will then show up in that the correlation between different traits measured with the same method will be higher than the different traits measured across methods, that is, $\min(r_{A_1,A_2}, r_{B_1,B_2}) > \max(r_{A_1,B_2}, r_{A_2,B_1})$. If there is no convergent validity, conclusions could not be drawn from the pattern. If all four measures represent different constructs, the method variance pattern could occur merely because the constructs measured by A_1 and A_2 , or B_1 and B_2 , happen to be more highly related than the cross-method constructs, independent of method variance.

Table 1 contains a multitrait-multimethod matrix from the developmental work with the Job Satisfaction Survey (JSS; Spector, 1985). This scale measures nine facets of job satisfaction, using a summated rating scale format. Its convergent and discriminant validities were investigated by administering it with the Job Descriptive Index (JDI; Smith, Kendall, & Hulin, 1969), which is considered the most carefully developed satisfaction instrument available (e.g., Vroom, 1964, p. 100). The JDI uses an adjective checklist, which is considered to be a different method. Both scales were administered to a sample of 102 municipal employees. The multitrait-multimethod analysis of the five common scales is presented here.

As can be seen in Table 1, the scales demonstrate convergent validities in that the validity coefficients in bold face are relatively large in magnitude and larger than values in their own row and column. Thus, the determination of method variance can proceed. Correlations among different traits within methods can be found in the monomethod, heterotrait triangles (top and bottom-right triangles). Correlations among different traits

Table 1
Multitrait–Multimethod Matrix for Job Descriptive Index and Job Satisfaction Survey Subscales

Subscale	1	2	3	4	5	6	7	8	9	10
Job Descriptive Index										
1. Work	—									
2. Pay	27	—								
3. Promotion	47	25	—							
4. Supervision	31	23	31	—						
5. Coworkers	37	30	37	28	—					
Job Satisfaction Survey										
6. Work	66	24	32	24	23	—				
7. Pay	33	62	51	34	30	29	—			
8. Promotion	34	31	77	27	34	20	61	—		
9. Supervision	25	27	26	80	24	22	34	28	—	
10. Coworkers	32	18	30	26	61	25	20	25	30	—

Note. $N = 102$, $r > .19$ for $p < .05$. From “Measurement of Human Service Staff Satisfaction: Development of the Job Satisfaction Survey” by P. E. Spector, 1985, *American Journal of Community Psychology*, 13, p. 701. Copyright 1985 by the Plenum Publishing Corporation. Adapted by permission.

across methods are found in the heterotrait, heteromethod triangles (remaining bottom triangles). The mean correlation for the 20 monomethod entries was .31, and for the 20 heteromethod entries was .29. These correlations were not significantly different from one another ($z = .16$, $p > .05$). Each pair of corresponding monomethod and heteromethod correlations was compared. Of the 40 pairs, the monomethod was larger than the corresponding heteromethod 24 times. These frequencies were not significantly different from one another, $\chi^2(1, N = 102) = 1.6$, $p > .05$. Not only was the mean correlation nonsignificantly larger in the monomethod case, the difference in magnitude was trivial.

This analysis was repeated for nine additional published studies. Four other studies were located, but their lack of convergent validities obviated method variance analysis. Some of these analyses involved comparing questionnaire with non-questionnaire methods. The lack of convergence raises questions about the validity of some of these measures. Six of the studies involved job satisfaction; two, job characteristics; and one, burnout. All but one of the job satisfaction studies involved comparison of different formats or instruments to measure the constructs of interest. One compared a questionnaire with an interview format, and one compared different language versions and different formats.

As can be seen in Table 2, which summarizes the results from all 10 analyses, in all but 1 case (Dunham, Smith, & Blackburn, 1977), method variance was nonsignificant and of extremely small magnitude. Although it is not clear why the Dunham et al. study produced method variance, the administration of multiple repeated measures may have led subjects to be careless and adopt response sets or styles that produced method variance.

One additional piece of evidence for method variance comes from Harvey, Billings, and Nilan’s (1985) confirmatory factor analysis of the Job Diagnostic Survey (JDS; Hackman & Oldham, 1975). Harvey et al. noted that each of seven subscales

of the JDS contains three items, each of a somewhat different format. These investigators placed data from a sample of employees into a multitrait–multimethod matrix, and then conducted covariance structure analysis (LISREL; Jöreskog & Sörbom, 1979). The researchers fitted several models and found that a model containing two latent factors of method variance best fitted the data. This model, however, still contained the seven expected subscales built into the instrument.

The results of these multitrait–multimethod analyses showed

Table 2
Summary of Multitrait–Multimethod Analysis of Method Variance

Study	<i>n</i>	Mean <i>r</i>		<i>z</i>
		Mono-method	Hetero-method	
Job satisfaction				
Alderfer, 1967	302	.07	.08	0.07
Dunham et al., 1977	622	.49	.31	3.78
Gillet & Schwab, 1975	273	.29	.23	0.68
Johnson et al., 1982	100	.40	.31	0.72
McCabe et al., 1980	82	.47	.44	0.24
Soutar & Weaver, 1982	242	.37	.34	0.37
Spector, 1985	102	.31	.29	0.16
Job characteristics				
Pierce & Dunham, 1978	155	.46	.39	0.75
Sims et al., 1976	941	.26	.22	0.92
Burnout				
Meier, 1984	320	.11	.15	0.51

Note. All correlations were heterotrait.

little evidence for method variance. They represent overall tests that do not reflect the actual sources of method variance. Why 1 of the 10 matrices located showed evidence for method variance with measures that were resistant in other studies is not clear. It may be that method variance does occur under certain conditions that may lead respondents to be careless. The strongest evidence for method variance came from the Harvey et al. (1985) study, which analyzed individual items from the JDS. Perhaps method variance may occur on the individual item level, although it does not occur on the subscale level. Additional research should be conducted to test this possibility.

Social Desirability

Six studies were found that related social desirability to variables of interest. There have been four attempts to correlate social desirability with job satisfaction. Lewis (1977) and Gansster et al. (1983) reported little relation between Crown and Marlowe's (1964) SD scale and the five subscales of the JDI. Correlations were all nonsignificant, ranging from $-.14$ to $.08$ across the two studies. Spector (1984) correlated the SD scale with the nine subscales of the JSS. Correlations ranged from $-.19$ to $.05$ and were all nonsignificant. Arnold et al. (1985) reported a significant $-.25$ correlation between SD and a single item measure of global satisfaction.

A single study was found reporting correlations between social desirability and perceived job characteristics. Brief and Aldag (1978) found one of six correlations significant ($r = .19$) between SD and the subscales of the Job Characteristics Inventory (Sims, Szilagyi, & Keller, 1976). Ganster et al. (1983) reported significant correlations between SD and role ambiguity and conflict ($r_s = -.27, -.28$, respectively). Using an expanded sample, however, Rosenkrantz, Luthans, and Hennessey (1983) found a significant correlation only for role ambiguity.

In addition to reporting correlations, Ganster et al. tested for spuriousness, suppression, and moderation in relations among measures of perceived supervisory behavior, job satisfaction, role stress, and personality. These researchers performed 73 tests for each type of effect and found no more than 4 of each, exactly what would be expected by chance at their $p < .05$ alpha level. In addition, the significant effects were extremely small.

All of these results suggest that social desirability is at best only weakly related to the measures of interest here. Furthermore, where there is some relation, it is so small that it produces little biasing effect. Thus, it would seem that social desirability is not a source of method bias in the measurement of affect and of perceptions of organizations.

Acquiescence

Fewer studies have been concerned with the biasing effect of acquiescence. Smith et al. (1969) were concerned with the problem during the development of the JDI. Although they found some small relations between measures of acquiescence and JDI scores, acquiescence had little or no effect on the convergent validities of the subscales. Brief and Aldag (1978) correlated scores on the Job Characteristics Inventory with acquiescence. Again they found one of six significant ($r = -.17$).

Data from Spector's (1985) JSS development provides another test for acquiescence response bias in job satisfaction. Be-

Table 3

Frequency of Acquiescent and Nonacquiescent Subjects on the Job Satisfaction Survey

Sum	Frequency	%
2	0	0
3	1	0.03
4	3	0.10
5	6	0.20
6	171	5.44
7	1,361	43.30
8	1,400	44.54
9	189	6.01
10	12	0.38
11	0	0
12	0	0
Total	3,143	

Note. Sums were generated by calculating separate mean agreement scores for positively and negatively worded items, and then summing the two scores. Sums less than 5 or greater than 9 indicate bias.

cause about one half of the items of this scale were positively and about one half negatively worded, acquiescence could be detected by comparing responses to both types of items (Winkler, Kanouse, & Ware, 1982). Subjects who tended to either agree or disagree with all of the items would show similar responses to each set of items. Separate scores were calculated for each item type and were divided by the number of items to yield an average score per item. Acquiescence was operationally defined as scores on both item types that suggested a tendency to agree or disagree; that is, scores above 4.5 or below 2.5 on a 6-point agree-versus-disagree scale. Responses to both item types were summed, and could range from 2 to 12. Table 3 summarizes the results from 3,143 respondents in the developmental sample. As can be seen, only 16 of 3,143 or .51% of subjects displayed the bias.

These results, as with social desirability, suggest that the biasing effect of acquiescence is quite small or nonexistent. Not only is acquiescence only minimally correlated with the measures of interest, it is demonstrated in very few subjects when responding to job satisfaction scales.

Spurious Relation Between Satisfaction and Absenteeism

Many researchers are concerned about spurious relations between affective variables and self-reports of even objective events such as absenteeism. Data on this question are available from an unpublished study by the author. Data on absenteeism were collected from both self-reports and organizational records on 66 employees of a human service organization. Both absenteeism measures were compared and correlated with job satisfaction. The correlation between the self-report and records measures of absence was moderately high ($r = .58$). Their means, however, were amazingly close (2.09 vs. 2.14 days over 3 months, respectively), which were not significantly different from one another, $t(65) = .24, p > .05$.

In addition to absenteeism, self-report data were collected on 20 variables, including 10 measures of satisfaction from the JSS, 8 measures of perceived job characteristics from the JDS,

organizational commitment, and perceived participation in decision making. Correlations were calculated between each absence measure and all 20 affective and perceptual variables. Each corresponding pair of correlations was compared. In not 1 of 20 comparisons was there a statistically significant difference using a *t* test for dependent correlation comparisons. The mean across 20 correlations was $-.13$ for self-report and $-.12$ for records.

The high degree of correspondence between the two measures of absenteeism suggests that there is no bias in their measurement. The less than perfect correlation between them probably can be attributed to random error. The lack of significant differences between correlations of affective and perceptual variables with the two measures of absenteeism is further evidence against the method variance problem. Correlations were not spuriously increased by collecting absence data from self-reports.

Conclusions

Method variance is a frequent criticism of research involving affective and perceptual variables in the I/O area. The data and research results summarized here suggest that the problem may in fact be mythical. Only 1 of 10 multitrait-multimethod analyses found evidence for method variance on the subscale level, and a single study found it on the item level. Correlations of bias measures with instruments designed to measure constructs of interest tend to be very small and rarely statistically significant. Their effects on relations among variables are minor at best.

One must be cautious in generalizing the results here to alternative measures of the same constructs. Most of the studies involved well-validated instruments with reasonably sound psychometric properties. Two of the studies in which evidence was found for method variance (Arnold et al., 1985; Harvey et al., 1985) found it at the level of single item measures. Method variance might well be more of a problem with single items or poorly designed scales. Researchers should explore this possibility in the future.

That method variance did not occur with the constructs of interest here does not mean that it does not exist with measures of other constructs. The measurement of personality has long been troubled by bias and distortion by respondents (Nunnally, 1978, pp. 596–606). Winkler et al. (1982) found that 5% of their subjects displayed acquiescence on their health attitude scale. Schriesheim and his colleagues (Schriesheim, 1981; Schriesheim, Kinicki, & Schriesheim, 1979) studied reports of supervisor behavior and found leniency problems, some of which may have been due to method variance. In the I/O area, many studies have been concerned with bias in the rating of job performance. Thus, the conclusions here should be generalized to other constructs very cautiously.

The conclusion that method variance is not a troublesome artifact in the measurement of several important organizational variables does not mean that their instruments are necessarily valid. Certainly, researchers who conduct research with self-report measures are sometimes troubled that the variables are too highly intercorrelated. One of the most important questions to address with this type of research concerns the reasons for high intercorrelations. For example, the current author has com-

pleted a study concerning job stress in clerical workers. High intercorrelations were observed among perceptions of such environmental conditions as work load and autonomy, job satisfaction, anxiety, experienced frustration, and somatic symptoms. If method variance cannot account for these relations, where does the explanation lie?

It is certainly possible that all of the instruments were valid measures of their intended constructs, and observed correlations reflected real relations among environmental conditions, affect, and health outcomes. Before accepting this explanation, however, one must be able to rule out equally plausible, alternative explanations. Several have been suggested and evidence exists to support their possibility. Staw (1975) suggested that individuals respond to self-report measures in ways that maintain consistency among attitudes, perceptions, and attributions. Attributions held by respondents cause them to give responses that are intercorrelated. Thus, if a respondent is satisfied with his or her job and believes that satisfying jobs in general are higher in variety, he or she will perceive the job as high in variety to achieve consistency (B. M. Staw, personal communication, January 5, 1986). Alternately, Staw and Ross (1985) provided evidence that job satisfaction may be a dispositional trait. Responses to various instruments may intercorrelate to the extent that they are in part caused by common underlying traits of the individual. Finally, James and Jones (1980), and James and Tetrick (1986), used complex modeling to suggest that self-reports of job and organizational characteristics may be an effect as well as a cause of job satisfaction. These explanations and others must be ruled out before one can draw the conclusion from self-reports that environmental and job conditions have particular effects on individuals. Until the relations among these self-report responses are understood, it will be difficult to draw definitive conclusions from the vast organizational literature that is heavily dependent on such measures. Unfortunately, most studies use single data source, cross-sectional designs, which do not allow investigation of these alternative explanations, let alone determination of method variance problems.

Future research should be directed specifically at determining the conditions under which method variance occurs and the extent to which it distorts results of studies. Of particular concern are the types of constructs that are subject to the problem, and the characteristics of instruments that reduce it. Studies using multiple sources of data can allow for multitrait-multimethod analyses and comparisons of relations of different sources with specific criteria. Instruments and constructs that may be resistant are suggested in the current study. How far this resistance generalizes is open to question.

References

- Aldag, R. J., Barr, S. H., & Brief, A. P. (1981). Measurement of perceived task characteristics, *Psychological Bulletin*, 90, 415–431.
- Alderfer, C. (1967). Convergent and discriminant validation of satisfaction and desire measures by interviews and questionnaires. *Journal of Applied Psychology*, 51, 509–520.
- Arnold, H. J., Feldman, D. C., & Purbhoo, M. (1985). The role of social-desirability response bias in turnover research. *Academy of Management Journal*, 28, 955–966.
- Brief, A. P., & Aldag, R. J. (1978). The Job Characteristic Inventory: An examination. *Academy of Management Journal*, 21, 659–670.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant

- validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.
- Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, 10, 3-31.
- Crowne, D. P., & Marlowe, D. (1964). *The approval motive*. New York: Wiley.
- Dunham, R. B., Smith, F. J., & Blackburn, R. S. (1977). Validation of the Index of Organizational Reactions with the JDI, the MSQ, and the Faces scales. *Academy of Management Journal*, 20, 420-432.
- Ganster, D. C., Hennessey, H. W., & Luthans, F. (1983). Social desirability response effects: Three alternative models. *Academy of Management Journal*, 26, 321-331.
- Gillet, B., & Schwab, D. P. (1975). Convergent and discriminant validities of corresponding Job Descriptive Index and Minnesota Satisfaction Questionnaire scales. *Journal of Applied Psychology*, 60, 313-317.
- Hackman, J. R., & Oldham, G. R. (1975). Development of the Job Diagnostic Survey. *Journal of Applied Psychology*, 60, 159-170.
- Harvey, R. J., Billings, R. S., & Nilan, K. J. (1985). Confirmatory factor analysis of the Job Diagnostic Survey: Good news and bad news. *Journal of Applied Psychology*, 70, 461-468.
- Husek, T. R. (1961). Acquiescence as a response set and as a personality characteristic. *Educational and Psychological Measurement*, 21, 295-307.
- James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, 69, 85-98.
- James, L. R., & Jones, A. P. (1980). Perceived job characteristics and job satisfaction: An examination of reciprocal causation. *Personnel Psychology*, 33, 97-135.
- James, L. R., & Tetrick, L. E. (1986). Confirmatory analytic tests of three causal models relating job perceptions to job satisfaction. *Journal of Applied Psychology*, 71, 77-82.
- Johnson, S. M., Smith, P. C., & Tucker, S. M. (1982). Response format of the Job Descriptive Index: Assessment of reliability and validity by the multitrait-multimethod matrix. *Journal of Applied Psychology*, 67, 500-505.
- Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Books.
- Lewis, B. A. B. (1977). *Effects of social desirability response bias on the measurement of job satisfaction*. Unpublished master's thesis, University of South Florida, Tampa.
- Marsh, H. W. (1983). Multitrait-multimethod analysis: Distinguishing between items and traits. *Educational and Psychological Measurement*, 43, 351-358.
- McCabe, D. J., Dalessio, A., Briga, J., & Sasaki, J. (1980). The convergent and discriminant validities between the IOR and the JDI: English and Spanish forms. *Academy of Management Journal*, 23, 778-786.
- Meier, S. T. (1984). The construct validity of burnout. *Journal of Occupational Psychology*, 57, 211-219.
- Michaels, C. E., & Spector, P. E. (1982). Causes of employee turnover: A test of the Mobley, Griffeth, Hand, and Meglino model. *Journal of Applied Psychology*, 67, 53-59.
- Mowday, R. T., Steers, R. M., & Porter, L. W. (1979). The measurement of organizational commitment. *Journal of Vocational Behavior*, 14, 224-247.
- Nicholls, J. G., Licht, B. G., & Pearl, R. A. (1982). Some dangers of using personality questionnaires to study personality. *Psychological Bulletin*, 92, 572-580.
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed). New York: McGraw-Hill.
- Pierce, J. L., & Dunham, R. B. (1978). The measurement of perceived job characteristics: The Job Diagnostic Survey versus the Job Characteristics Inventory. *Academy of Management Journal*, 21, 123-128.
- Roberts, K. H., & Glick, W. (1981). The job characteristics approach to task design: A critical review. *Journal of Applied Psychology*, 66, 193-217.
- Rosenkrantz, S. A., Luthans, F., & Hennessey, H. W. (1983). Role conflict and ambiguity scales: An evaluation of psychometric properties and the role of social desirability response bias. *Educational and Psychological Measurement*, 43, 957-970.
- Schriesheim, C. A. (1981). The effect of grouping or randomizing items on leniency response bias. *Educational and Psychological Measurement*, 41, 401-411.
- Schriesheim, C. A., Kinicki, A. J., & Schriesheim, J. F. (1979). The effect of leniency on leader behavior descriptions. *Organizational Behavior and Human Performance*, 23, 1-29.
- Sims, H. P., Jr., Szilagyi, A. D., & Keller, R. T. (1976). The measurement of job characteristics. *Academy of Management Journal*, 19, 195-212.
- Smith, P. C., Kendall, L. M., & Hulin, C. L. (1969). *The measurement of satisfaction in work and retirement*. Chicago: Rand-McNally.
- Soutar, G. N., & Weaver, J. R. (1982). The measurement of shop-floor job satisfaction: The convergent and discriminant validity of the Worker Opinion Survey. *Journal of Occupational Psychology*, 55, 27-33.
- Spector, P. E. (1984). *Development of the Work Locus of Control Scale*. Unpublished manuscript, University of South Florida, Tampa.
- Spector, P. E. (1985). Measurement of human service staff satisfaction: Development of the Job Satisfaction Survey. *American Journal of Community Psychology*, 13, 693-713.
- Staw, B. M. (1975). Attribution of the "causes" of performance: A new alternative interpretation of cross-sectional research on organizations. *Organizational Behavior and Human Performance*, 13, 414-432.
- Staw, B. M., & Ross, J. (1985). Stability in the midst of change: A dispositional approach to job attitudes. *Journal of Applied Psychology*, 70, 469-480.
- Vroom, V. H. (1964). *Work and motivation*. New York: Wiley.
- Winkler, J. D., Kanouse, D. E., & Ware, J. E., Jr. (1982). Controlling for acquiescence response set in scale development. *Journal of Applied Psychology*, 67, 555-561.

Received September 8, 1986

Revision received December 17, 1986

Accepted November 22, 1986 ■

Situational Leadership Theory: An Examination of a Prescriptive Theory

Robert P. Vecchio
University of Notre Dame

In a study of 303 teachers representing 14 high schools, measures were taken of supervisory style (consideration and initiating structure), follower maturity, performance, satisfaction with supervision, and quality of leader-member exchange. A variety of statistical tests were conducted to test the prescriptions for effective supervision contained in Situational Leadership Theory (Hersey & Blanchard, 1982). Results suggest that the theory may hold only for certain types of employees. Specifically, the results imply that more recently hired employees may need and appreciate greater task structuring from their superior. These results have implications for reinterpreting the theory and examining it within the "substitutes for leadership" perspective (Kerr & Jermier, 1978).

Hersey and Blanchard's (1982) Situational Leadership Theory (SLT) embodies one of the more widely known and, at the same time, least researched views of managerial effectiveness. Throughout this article, the term *managerial* will be used interchangeably with the term *leadership* as SLT focuses on the effectiveness of nominal heads rather than on emergent or incremental forms of power and influence (i.e., leadership per se). As noted by Graeff (1983), the theory is often cited in academically oriented management textbooks. However, we can offer little advice to our students, or to practicing managers, on the utility of the theory. Before we can endorse or critique the theory to our constituencies, a rigorous test of the theory's propositions is, of course, required. In this article, the origins and central elements of the theory, the available evidence of the theory's validity, and the requirements for a rigorous test of the theory's propositions are considered.

Origins and Elements of Situational Leadership Theory

Situational Leadership Theory developed from the writings of Reddin (1967). Reddin's 3-Dimensional Management Style Theory posits the importance of a manager's relationship orientation and task orientation in conjunction with effectiveness. From the interplay of these dimensions, Reddin proposed a typology of management styles (e.g., the autocrat, the missionary, the deserter). Although Reddin suggested that his framework explained effectiveness as a function of matching style to situation, his approach did not identify specific situational attributes that could be explicitly incorporated into a predictive scheme.

Building on Reddin's (1967) suggestion that leader or manager effectiveness varies according to style, Hersey and Blanch-

ard (1969) proposed a life-cycle theory of leadership. According to life-cycle theory, degrees of task orientation and relationship orientation must be examined in conjunction with the dimension of follower maturity to account for leader effectiveness. The central precept of life-cycle theory (1969, p. 29) is that as the level of follower maturity increases, effective leader behavior will involve less structuring (task orientation) and less socio-emotional support (relationship orientation). However, the decline in need for both of these leader behaviors is not straightforward. During the early stages of an employee's tenure, a low level of relationship orientation coupled with high task orientation is considered to be ideal. As an employee (or group of employees of roughly equal maturity) gains in maturity, the need for supervisory social-emotional support increases, while the need for structuring declines. Beyond a certain level of maturity, the need for both social-emotional support and structuring declines. At the highest levels of employee maturity, supervisory task and social behaviors become superfluous to effective employee performance.

In a popular text (evidenced by its being in its 4th edition), Hersey and Blanchard (1982) attempted to provide still greater precision to these precepts. They suggested that follower maturity can be broken into benchmark categories of high, moderate, and low, and that appropriate leader style can be summarized in terms of a leader primarily telling, selling, participating, or delegating in relations with subordinates. This most recent statement of the Hersey-Blanchard SLT model (1982, p. 152) is summarized in Figure 1.

Evidence of the Model's Validity

At a purely theoretical level, SLT has been suggested as having a good deal of overlap with other popular views of leader and group behavior. In a comparison of their views with those offered by other perspectives, Hersey and Blanchard (1982, chap. 13) achieved a synthesis of their concepts with those contained in McGregor's (1960) Theory X and Y, Argyris's (1957) maturity-immaturity continuum, Likert's (1967) management

The author expresses his appreciation to Cindy Harrison, who helped to ensure the cooperation of all participants.

Correspondence concerning this article should be addressed to Robert P. Vecchio, Department of Management, University of Notre Dame, Notre Dame, Indiana 46556.

STYLE OF LEADER

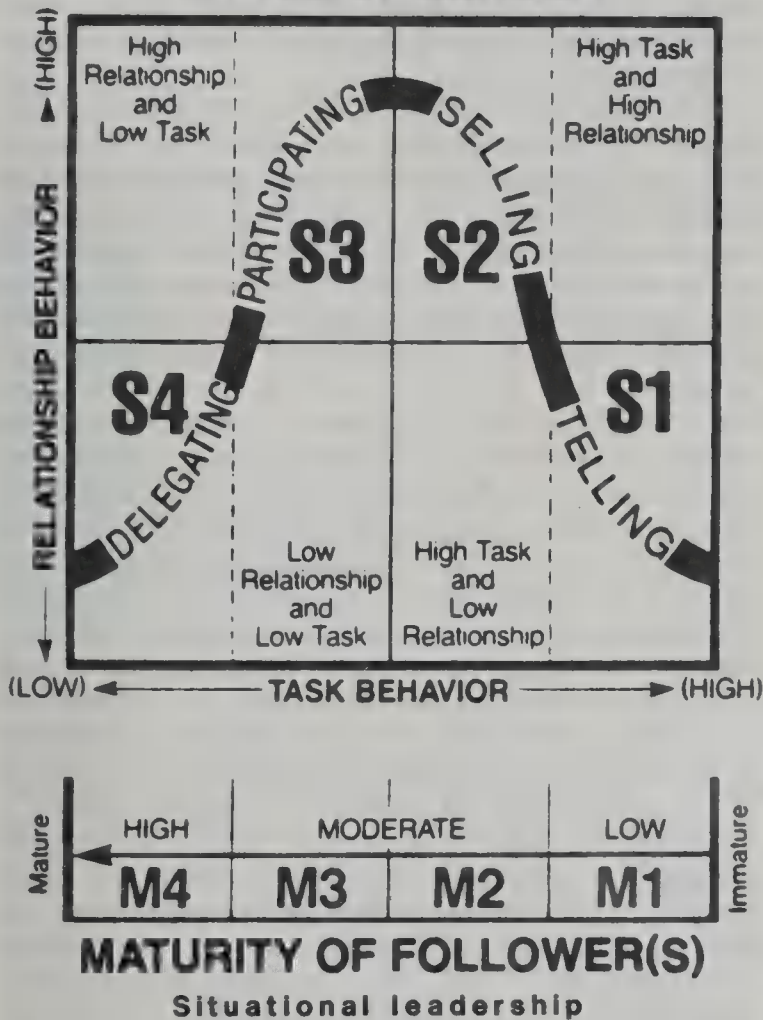


Figure 1. The situational leadership model. (From *Managing Organizational Behavior* [p. 152] by P. Hersey and K. Blanchard, 1982, Englewood Cliffs, NJ: Prentice Hall. Copyright 1982 by Prentice-Hall, Inc. Reprinted by permission.)

systems, Maslow's (1954) need hierarchy, Herzberg's (1966) two-factor theory, McClelland's (1961) achievement theory, Schein's (1970) assumptions of human nature, transactional analysis (Berne, 1964; Harris 1969), French and Raven's (1959) power bases, parent effectiveness training concepts (Gordon, 1970), Greiner's (1972) phases of organizational growth, Lewin's (1947) views of achieving behavioral change, behavior modification (Skinner, 1953), and force field analysis (Lewin, 1947). Although one can cynically argue that this high degree of overlap suggests that SLT is not saying much that is new or original, it can also be contended that many of the above theories can also be shown to contain a high degree of overlap. More positively, one can argue that SLT is focusing on critical features of behavior that have been previously identified.

The fact that SLT can be shown to overlap to varying degrees with other theories is not in itself sufficient evidence of SLT's validity. This point is of clear importance when one considers that many of the previously cited theories have not achieved a high degree of empirical support (e.g., Herzberg's two-factor theory), despite a fair amount of attention in the academic literature. In terms of internal theoretical coherence, Graeff (1983)

has provided the most comprehensive critique of the theory. In his review of SLT, Graeff argued that the theory may actually have been derived from a passage in an article by Korman (1966), in which the suggestion of curvilinear relationships between dimensions of leader behavior and other variables was taken to mean that a curvilinear relationship may exist between dimensions of leader behavior. In addition, Graeff (1983) suggested that the manner in which components of dimensions in SLT are combined and the manner of graphic presentation of a four-dimensional model (task orientation, relationship orientation, follower maturity, and effectiveness) in only two dimensions are critical problems for the theory. Also, he has suggested that the popularly advocated measurement device for studying leader behaviors (the LEAD instrument) possesses unknown psychometric qualities. In the theory's favor, Graeff argued, however, that SLT correctly focuses on issues of leader flexibility and the importance of subordinate attributes as the key situational determinant of appropriate leader behavior.

At an empirical level, the theory has received little attention. One of the earliest published studies devoted to SLT concepts focused on the development of a measure of follower maturity, yet it did not use the measure in a test of the model (Moore, 1976). In a study that approximated a test of the theory, Hambleton and Gumpert (1982) asked managers to select at random 4 of their subordinates to complete a survey instrument. For 65 participating managers (of 159 who were contacted), manager ratings of subordinate maturity were coded in conjunction with manager self-assessments of leadership style (high vs. low task and relationship orientation). From this coding, matches and mismatches were identified. Matches occurred in only 29% of the cases. A comparison of mean performance ratings—given for each employee by the managers—for the matches and mismatches revealed that the matches received somewhat higher mean evaluations ($t = 6.47, p < .01$).

Although the findings of Hambleton and Gumpert (1982) are the only available supportive evidence for the model, they raise several concerns. First, the sample suffered severe attrition (i.e., less than half of the managers provided data). Second, the managers provided self-assessments of their own style. Such self-assessments of leader behavior are not regarded as being highly accurate (Schriesheim & Kerr, 1974). In addition, these assessments were taken on a version of the LEAD instrument, rather than on a more widely studied and accepted measure of leader behavior. Last, the respondents were highly cognizant of SLT precepts (as evidenced by their having been asked to rate their knowledge of SLT and to assess the extent to which they used SLT in their work). This awareness of SLT principles on the part of the participants may have induced some respondents to attempt to complete their surveys in conformity with the theory (i.e., in order to appear to be applying their knowledge of the theory). Also, all of the respondents reported at least fair knowledge of and some use of SLT.

The most recently reported study of SLT (Blank, Weitzel, & Green, 1986) involved 27 hall directors and 353 resident advisors (subordinates) at two large universities. Respondents completed the Leader Behavior Description Questionnaire (LBDQ-XII; Stogdill & Coons, 1957) and a measure of maturity. Directors provided performance ratings of resident advisors, and

each resident advisor completed subscales of the Job Description Index (JDI) satisfaction measure (Smith, Kendall, & Hulin, 1969). In their analysis of these data, Blank, Weitzel, and Green did not report the results of matching subordinate maturity and leader behaviors to predict subordinate performance and satisfaction, but instead examined interactions between maturity and each of the two leader behavior dimensions (consideration and structuring) in an attempt to predict subordinate performance and satisfaction. In general, their search for two-way interactions did not reveal support for the theory. However, it should be noted that the theory predicts a three-way interaction, and not separate two-way interactions, among the key variables (i.e., the interaction of maturity with structuring should not be examined independently of consideration, but jointly).

In summary, investigations of the theoretical and empirical robustness of SLT have been rare. Although the theory contains strong intuitive appeal, the veracity of the theory has not been assessed via a rigorous empirical test.

Issues Surrounding a Test of the Situational Leadership Theory

In order to test SLT, several issues must be addressed that relate to the clarity of the theory's prediction. One major issue surrounds the unit of analysis for SLT: the individual versus the group. Although the theory is often stated in terms of group maturity, there are also many references to individual maturity as well. For example, in their definition of maturity, Hersey and Blanchard (1982, p. 151) stated that an "individual or a group" is their focus. They also recognized that when one relates to an entire group (e.g., a teacher speaking to a class of students), it is the maturity level of the group that is important. However, when one deals with an employee in a one-to-one setting (e.g., a teacher speaking with a single student), the maturity level of the individual is most important. This recognition of the need for leaders to behave differently with individual group members than when they relate to an entire group is an important statement. Much of the recent research in the area of leadership has, in fact, focused on the issue of universal versus differential leadership style (cf. research on the Vertical Dyad Linkage Model; Liden & Graen, 1980). That the dynamics of SLT are presumed to operate at both levels is an important feature of the SLT framework. In the context of a test of SLT, it is necessary to specify and be consistent in studying leadership phenomena at a given level (and not across levels) of analysis. It seems likely—in light of the preponderance of research at the individual level and the suspicion that group processes may mask individual process—that SLT will be most robust at the individual level of social dynamics (i.e., leadership behavior that is in accord with the prescriptions of SLT will be more effective when it is targeted to a given individual's level of maturity).

An extended issue that is beyond the present investigation is whether individual maturity interacts with group maturity in determining leader effectiveness. It is easy to envision a situation in which a subordinate is significantly more mature than his or her peers in a given position, yet the leader's behaviors (if they are often displayed for the benefit of the group) may be

grossly inappropriate. In such settings, the incongruent subordinate will likely be dissatisfied with the leader's directions and may be resentful of the limits that his peers indirectly set via the leader's actions.

Perhaps the least clear feature of SLT surrounds the definition of effectiveness. In their book, Hersey and Blanchard (1982, pp. 96–99) define effectiveness and ineffectiveness as occurring when a leader's style is appropriate and inappropriate, respectively. Although they recognize that effectiveness can be viewed as a continuum, they do not acknowledge the multifaceted nature of the concept. Also, the definition of effectiveness in terms of appropriateness of leader style is somewhat circular in its use of logic. In order to test SLT, it is useful to define effectiveness in broader, more traditional terms. For example, effectiveness can be defined in terms of output, cost reduction, enhancement of employee motivation, morale, and so forth. This restatement of effectiveness implies that SLT should be restated so that effectiveness is a possible outcome of appropriateness of leader behavior. The use of the word *possible* is important in the foregoing sentence because an appropriate combination of leader style (in terms of SLT's prescription) may still not enhance subordinate behavior and attitudes. As noted by Kerr and Jermier (1978), situational attributes can offset and substitute for leader behaviors. In a sense, subordinate maturity in SLT is a substitute for leadership, in that subordinates of higher maturity need less attention or direction from leaders. Increases in subordinate self-sufficiency (maturity), which likely result from relevant work experience and training, can make leader behaviors increasingly irrelevant to subordinate performance and morale. Therefore, SLT's prediction that highly mature employees require a low-structure–low-consideration style of supervision may be partially misstated. It may be more correct to say that supervisory style is comparatively more irrelevant, in terms of its impact on highly mature subordinates. In short, the conduct of highly mature subordinates may simply be less predictable than that of other employees, from supervisory attributes.

A further issue centers on how to test in a statistical sense, the predictive accuracy of SLT. At its heart, the theory forecasts a three-way interaction of leader consideration, leader structuring, and subordinate maturity. If one imagines the form of this hypothesized interaction by trying to graph the hyperplane that is predicted, it becomes apparent that the predicted interaction does not satisfy the statistical assumption of homoscedasticity. That is to say, the regression-based assumption of equal variance around the regression plane does not hold, by definition, for SLT as the predictions of superior performance only hold for specific points in the multidimensional space. In all other locations in the space, the data are free to vary. Therefore, the theory only makes predictions for specific combinations of variables. For all other combinations, the theory is silent. In essence, a test of the theory that uses the statistical technique of multiple regression may lead, erroneously, to the conclusion that the theory is incorrect for a given data set (i.e., a Type II error). To more fairly test the theory, it would be worthwhile to examine the predictions in the manner proposed by the theory's developers (i.e., to compare the effectiveness of leaders whose styles are "appropriate" for given settings to leaders whose styles are pre-

dicted to be "inappropriate" for the same settings). Although one may test whether a given data set violates the requirement of homoscedasticity, the results of such a test would not, of course, shed light on the validity of Situational Leadership Theory, *per se*.

The purpose of the present investigation was (a) to test SLT in a study that was designed to capture the critical variables proposed by the theory and (b) to explore SLT with analytic techniques that reflect traditional practice in organizational research (i.e., regression) as well as accommodate the need to study leadership as a situation-specific phenomena (i.e., subgrouping analysis).

Method

Subjects and Procedure

Subjects were 303 full-time high school teachers, who represented 14 high schools in a large midwestern city and provided data in response to a confidential survey. Because of the support of the head of the school district, cooperation was readily obtained from the principals of all 14 schools in the district. During monthly meetings with faculty, the principals distributed surveys to their teachers. At the meetings, time was devoted to the completion of the surveys. Both teachers and principals completed similar surveys, which focused on the behavior of the school principal (leader) and the individual teacher (subordinate). To ensure anonymity for the teachers, the surveys were coded with ID numbers. In addition, the completed surveys of both teachers and principals were collected at the meeting, placed in an envelope, and mailed directly to the author. With the exception of 34 teachers who did not attend their school's monthly meeting in January 1986, all of the principals and teachers responded to the survey.

Measures

Each teacher provided responses to the following scales: (a) JDI, satisfaction with supervision (Smith et al., 1969); (b) Leader-Member Exchange, quality of leader-member relationship (Liden & Graen, 1980); (c) LBDQ-XII, leader consideration (Stogdill & Coons, 1957); and (d) LBDQ-XII, leader initiating structure (Stogdill & Coons, 1957). Modified versions of the LBDQ-XII measures of leader behavior were used in place of the LEAD instrument because of the relative psychometric advantages of the LBDQ-XII (i.e., its reliability and construct validity have received more attention than the LEAD instrument, and it is a more widely accepted index of leader behavior than the LEAD instrument). In addition, the stems of the items in the LBDQ-XII used in this study were modified to incorporate an individualized format (cf. Vecchio & Gobdel, 1984): sample item, "My principal acts without consulting me." Furthermore, each teacher was asked to complete a follower maturity index that contained items related to both task-relevant (e.g., understanding of job requirements) and psychological (e.g., commitment) forms of maturity (Hambleton, Blanchard, & Hersey, 1977).

Principals provided ratings for each teacher on dimensions of follower maturity and performance. Maturity was assessed on items related to task-relevant and psychological maturity, whereas performance was assessed by summing ratings across dimensions of dependability, planning, know-how, present performance, and expected performance (Liden & Graen, 1980).

Analytic Techniques

The accuracy of the principles of SLT was examined with several statistical techniques. As will be shown, each technique represents ■

somewhat different phrasing of the central research question. The first technique used was hierarchical regression analysis, in which a three-way multiplicative interaction term was created (Maturity \times Consideration \times Structuring). This interaction term was entered into a regression equation following the inclusion of main effects and two-way interaction terms to determine whether the inclusion of the three-way interaction term appreciably increased the variance accounted for in the performance criterion. As was noted earlier, the use of multiple regression for testing SLT can be critiqued on the grounds that certain assumptions of regression cannot be met when the principles of SLT are, in fact, correct. Nonetheless, the robustness of regression techniques and our present uncertainty as to just what would result with the use of the technique warrant the exploratory use of multiple regression to test the theory.

A second approach to testing SLT involves the creation of subgroups of employees for whom the theory is expected to hold or not to hold. This requires the creation of subgroups based on the combination of consideration, structuring, and maturity. For employees whose situations are designated "matches," their mean performance should be superior to that of subordinates for whom the situations are "mismatches." This comparison of matches and mismatches represents an omnibus test of SLT in that it ignores differences within specific categories of maturity in favor of an overall test.

It can be argued, however, that an omnibus test is not the best possible device for assessing SLT. If the distribution of cases is not uniform across categories, then the results of the omnibus test may be biased. For example, no (or very few) cases may exist for some combinations of maturity and leader behavior. The peculiarities of these distributions and the possible associated mean differences for the categories could, thereby, produce spurious results. A third, more direct, assessment of SLT involves the creation of several categories on the dimension of follower maturity. After creating these categories, comparisons could then be made within categories to determine whether subordinates who match on the leader dimensions are superior performers, relative to those who do not match on these dimensions. The need to first create categories of maturity is perhaps critical in that different levels of follower maturity are likely to be related to different levels of overall performance (although SLT does not directly address this critical issue). Therefore, the likely correlation of maturity with performance needs to be controlled by conducting comparisons within levels of maturity.

To be sure, only the third, partitioned, technique is the most defensible approach to assessing SLT. However, we presently know so little about SLT-related phenomena that all three techniques are reported here in the interest of completeness and in order to gain further understanding.

Results

Zero-order correlations among the predictors and criteria (see Table 1) reveal, first, that the variables are all, at least moderately, related to one another. Although this is to be expected from a survey-based study, it does not pose a serious obstacle to studying SLT principles in that SLT does not posit main effects, but interactions. Second, the existence of main effects makes it more difficult to identify complex forms of relationships because the main effects "steal" criterion variance (i.e., the present test of SLT can be viewed as being conservative in that it will be relatively difficult to uncover SLT effects). Table 1 also displays the internal consistency coefficients for the measures of interest. All of these coefficients are of reasonable magnitude (ranging from .82 to .94). Evidence that the maturity measure at least partially taps work-relevant experience was ob-

Table 1
Correlations Among Predictors and Criteria

Variable	M	SD	Range	1	2	3	4	5	6
1. Structuring ^a	16.89	4.20	5–25	82	—				
2. Consideration ^a	18.89	3.99	5–25	52	83	—			
3. Maturity ^b	40.26	6.34	15–48	26	35	93	—		
4. Performance ^b	29.62	4.87	10–35	27	31	85	94	—	
5. Leader–member quality ^a	21.46	4.58	7–28	60	79	34	35	91	—
6. Satisfaction with supervisor ^a	47.03	10.97	2–57	79	74	31	36	79	90

Note. Coefficient alphas are given on the primary diagonal. All correlations are significant beyond the .01 level. Decimal points are omitted.
^a Data provided by subordinates. ^b Data provided by supervisors.

tained by correlating individual maturity with self-reported years of teaching experience. The obtained correlation suggested that years of experience was related to the index ($r = .15$, $p < .01$). It should also be noted that the range of professional teaching experience for the sample was substantial (range = 1–31 years; $M = 20.5$ years).

The results of the hierarchical regression analyses are presented in Table 2 for the criteria of supervisor rating, leader–member exchange, and satisfaction with supervision. For each analysis, the three predictors of consideration, structuring, and maturity were entered simultaneously at the first step. Next, the two-way interaction terms were entered into the equation. Last, the three-way interaction term was included. The increment in R^2 at each step (i.e., in variance accounted for) was calculated and tested for significance (Cohen & Cohen, 1975). As the results in Table 2 indicate, none of the three criteria tested yielded support for a three-way interaction.

Omnibus tests of mean differences on the criteria were also conducted. For these analyses, the distribution of follower maturity was trichotomized into high, moderate, and low maturity by cutting at the values of 44 and 40. Three categories, rather than four, of maturity were created to ensure a sufficient number of cases for each subgroup. The predictions for the middle-range groups on maturity are identical (i.e., high consideration coupled with moderate structuring is prescribed). Next, the dimensions of consideration and structuring were trichotomized and dichotomized, respectively. Cuts on the structuring dimension were made at the values of 19 and 15. On the consideration dimension, the split was made at the value of 19. This resulted in a 3×2 cross tabulation, in accord with the SLT model. Em-

ployees whose values on maturity coincided with the prescribed levels of consideration and structuring were designated matches; all remaining employees were designated mismatches. It is, of course, predicted that mean values for the outcome variables will be higher for the matched group, relative to the mismatched group.

As Table 3 reveals, a large percentage of the employees were in the mismatched group (i.e., their situations were those for which the theory predicts lower effectiveness). This finding, in itself, is of some importance in that it suggests that the positive SLT prescriptions may have little relevance to a majority of employees (i.e., the natural occurrence of the preferred combinations may be fairly low). Alternatively, it can be argued that there is a great untapped potential or significant need for creating circumstances that the theory prescribes. To test this later point, mean differences were tested for significance for the matched versus mismatched groups. If the theory is correct, we can expect the matched group (albeit a smaller group) to have higher values on the outcome measures.

As the results reported in Table 3 indicate, the means were in the predicted direction for all three comparisons. In two of these comparisons, the means were significantly different (although the estimates of the effects' sizes were not substantial). The results suggest that employees who describe their superiors' behavior on the dimensions of consideration and structuring in accord with SLT prescriptions, given their specific level of maturity, tend to have somewhat higher performance ratings, and to report higher quality relationships with their supervisor, as well as greater satisfaction with their supervisor.

Although the omnibus tests provided evidence of the accu-

Table 2
Summary of Regression Analyses

Source	Performance		Leader–member quality		Satisfaction	
	R^2	ΔR^2	R^2	ΔR^2	R^2	ΔR^2
Consideration (C), structuring (S), maturity (M)	.743**		.686**		.639**	
$C \times S$, $C \times M$, $S \times M$.744**	.001	.697**	.011*	.641**	.012*
$C \times S \times M$.744**	.000	.697**	.000	.641**	.000

* $p < .05$. ** $p < .01$.

Table 3
Results of Omnibus Tests

Group	M	SD	n	F	p	Estimated effect size
Performance						
Match	30.4	3.91	50	1.61	.21	.002
Mismatch	29.5	5.03	252			
Leader-member quality						
Match	22.8	22.80	50	5.25	.02	.014
Mismatch	21.2	21.18	245			
Satisfaction with supervision						
Match	51.0	6.25	46	7.35	.01	.022
Mismatch	46.3	11.52	238			

racy of SLT, they do not tell us precisely where these differences are occurring in the framework. Also, the supportive results of the omnibus tests may have capitalized on unique, sample-specific differences that are correlated with uneven distributions of attributes. Therefore, sets of partitioned tests of the theory were conducted. These tests involved making mean difference comparisons within maturity groups. That is to say, matches and mismatches were designated within each of the three levels of maturity. The results of these comparisons (see Table 4) indicate that the mean differences were in the correct direction in six of nine comparisons. Of these six, four were significantly different. The effect sizes for these differences ranged from .032 to .160. In one instance, the mean difference was significant in the reverse direction of that which was hypothesized. Also, the results were more generally supportive in the low-maturity category, somewhat mixed in the moderate-maturity category, and generally nonsignificant in the high-maturity category. This suggests a very different picture of the results, in which the theory is only correct for low-maturity employees and less correct for more mature employees. The implication is that low consideration coupled with high structuring is a superior combination for low-maturity employees. For moderate- and especially high-maturity employees, it is not clear that the combinations prescribed by SLT are associated with superior outcomes. Furthermore, it is worth noting that the mismatched groups (Tables 3 and 4) were somewhat more variable than the matched groups. This is evidenced by the often larger standard deviations for the mismatched groups. This difference in variability confirms the earlier suggestion that the more stringent assumptions of parametric statistical tests (e.g., regression analysis) may not be satisfied by the data generated by—as well as the logic attendant to—tests of Situational Leadership Theory.

Discussion

The present study represents one of the first comprehensive tests of Situational Leadership Theory. As such, it is not possible to contrast the current findings with those obtained in other investigations. Therefore, the present evidence must be taken, for the moment, as providing the best available test of the theory's principles. In general, the present study provides partial

support for the theory in that the omnibus tests and several of the partitioned tests point to the theory being partially accurate in its prescriptions. The results of the regression analyses are perhaps suspect in that various assumptions of regression analysis are contrary to the essential features of SLT. Furthermore, the predictor variables that were used in the regression analysis were intercorrelated. This may be more than a common source effect problem because the predictor constructs may not be logically or empirically independent of one another.

The finding that SLT was most strongly supported in the low-maturity condition appears reasonable in that employees who are relatively lacking in task-relevant knowledge or commitment should require more structuring on the part of their supervisor. Displays of considerateness by superiors for low-maturity subordinates would be tantamount to sending improper signals to such subordinates. For subordinates of moderate maturity, it is not clear what style of supervision works best. The present

Table 4
Summary of Partitioned Tests

Group	M	SD	n	F	p	Estimated effect size
Low maturity						
Performance						
Match	27.3	3.50	21	4.68	.03	.032
Mismatch	25.0	4.54	92			
Leader-member quality						
Match	24.2	3.06	21	22.18	.00	.160
Mismatch	19.1	4.79	92			
Satisfaction with supervision						
Match	53.4	3.86	19	16.80	.00	.129
Mismatch	41.3	12.68	88			
Moderate maturity						
Performance						
Match	32.1	2.37	15	5.65	.02	.048
Mismatch	30.1	3.11	78			
Leader-member quality						
Match	19.6	3.20	15	4.18	.04	.034
Mismatch	22.1	4.54	75			
Satisfaction with supervision						
Match	47.0	7.51	13	<1		
Mismatch	48.6	9.84	72			
High maturity						
Performance						
Match	33.4	2.10	14	1.20	.28	.002
Mismatch	33.9	1.70	82			
Leader-member quality						
Match	24.1	2.81	14	1.65	.20	.007
Mismatch	22.7	3.89	80			
Satisfaction with supervision						
Match	51.4	6.15	14	<1		
Mismatch	49.6	9.66	78			

data suggest that performance is greater for these same employees if moderate structuring is combined with high consideration. However, the same sample provides evidence that the quality of leader-member relationships may be significantly lower when this particular combination of styles is reported.

For high-maturity employees, the theory appears to be unable to predict. Perhaps the theory needs to be rephrased to accommodate such high-maturity types. As it stands, the theory seems to suggest that highly mature employees can be relatively free from direction and do not need to receive "strokes" from their supervisors. Such a scenario is highly doubtful, as most employees probably appreciate supervisor considerateness and occasional signs of supervisor interest (as manifested by supervisor direction or structuring). In addition, the measurement of maturity poses a unique problem for testing Situational Leadership Theory. Self-reports of maturity are highly suspect. Similarly, peer ratings may largely reflect popularity rather than task orientation. Yet, the construct is so broad that a rating seems the most appropriate technique for addressing it. It is interesting that in the present sample, years of teaching was somewhat associated with supervisor ratings of maturity. However, experience can still be regarded as independent of supervisor ratings. This makes sense as long-tenured employees may be, on the average, more competent than are more recently hired individuals. Yet, the job-relevant maturity of long-tenured employees can still be quite variable (i.e., years of experience and job-relevant maturity are not likely to be highly correlated across situations). Of further interest, the correlation of teachers' self-ratings of maturity were significantly correlated with their superiors' ratings on the same instrument ($r = .28, p < .01$).

It is possible that SLT is not well suited to making predictions in any given job category, in that a full range of maturity and leader behaviors may not be manifested in one job classification. Perhaps SLT is better viewed as being prescriptive across job categories. In this view, high to low maturity represents various classes of jobs (e.g., professional to unskilled). For professional (high-maturity) jobs, supervisors should display relatively less consideration and less structuring. Professionals, as the reasoning goes, should be capable and desirous of greater self-direction. For unskilled (low-maturity) jobs, supervisors should provide significantly greater structuring and less consideration. Unskilled workers perhaps expect, and may prefer, greater direction and less social-emotional attention on the part of supervisors (cf. Vroom & Mann, 1960). This restatement of SLT suggests that the underlying principles of SLT may be valid but that the theory may be improperly conceptualized, such that the current focus is on maturity differences within jobs rather than across jobs.

An across-jobs perspective offers possibly greater ranges of maturity and, therefore, a greater likelihood of identifying systematic relationships. However, this across-jobs perspective requires a modification of the term *maturity*. As it is used in a within-job perspective, its meaning is fairly clear (i.e., employee level of task-relevant knowledge and commitment). In an across-job perspective, maturity may have to be replaced with a more level-appropriate concept such as normative expectations. This across-jobs view suggests that job prestige (or job quality) dictates specific norms for preferred styles of supervi-

sion. To be effective, a supervisor should be conscious of and responsive to these norms. The above viewpoint also can be seen as incorporating notions from Kerr and Jermier's (1979) suggestion of substitutes for leadership. In professional and highly skilled positions, experience and knowledge can make a supervisor's influence less important. In less skilled positions in which employees lack experience and acquired knowledge, supervisory influences can be of far greater importance. However, a testing of an across-jobs perspective should perhaps not be undertaken without further testing of the within-job view of Situational Leadership Theory. In addition, future tests of the theory should, to the degree possible, use independent measures of predictors.

In summary, the present study provided partial support for principles contained in Situational Leadership Theory. As is not uncommon in organizational research, an initial, seemingly simple research question yielded a complex set of results that only substantiated a portion of a set of propositions. Furthermore, the results underscored the somewhat disheartening observation that the approach taken in analyzing a theory determines, to an extent, the form and degree of support that is obtained for the propositions. The present results most strongly suggest that more recently hired employees may require greater structuring from their superior. Nonetheless, the present results are sufficiently intriguing so as to suggest that SLT be studied further, with an across-jobs perspective and with a recognition that high-maturity conditions may obviate the need for supervision, rather than specify a particular style of supervision.

References

- Argyris, C. (1957). *Personality and organization*. New York: Harper & Row.
- Berne, E. (1964). *Games people play*. New York: Grove Press.
- Blank, W., Wietzel, J., & Green, S. G. (1986). Situational leadership theory: A test of underlying assumptions. *Proceedings of the Academy of Management*, 384.
- Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- French, J. R. P., & Raven, B. (1959). The bases of social power. In D. Cartwright (Ed.), *Studies in social power* (pp. 150-167). Ann Arbor, MI: Institute for Social Research.
- Gordon, T. (1970). *Parent effectiveness training*. New York: P. H. Wyden.
- Graeff, C. L. (1983). The situational leadership theory: A critical view. *Academy of Management Review*, 8, 285-291.
- Greiner, L. E. (1972, July-August). Evolution and revolution as organizations grow. *Harvard Business Review*, 37-46.
- Hambleton, R. K., Blanchard, K. H., & Hersey, P. (1977). *Maturity Scale—Self-rating form*. San Diego, CA: Learning Resources Corporation.
- Hambleton, R. K., & Gumpert, R. (1982). The validity of Hersey and Blanchard's theory of leader effectiveness. *Group and Organization Studies*, 7, 225-242.
- Harris, T. (1969). *I'm OK—You're OK: A practical guide to transactional analysis*. New York: Harper & Row.
- Herzberg, F. (1966). *Work and the nature of man*. New York: World Publishing.
- Hersey, P., & Blanchard, K. (1969). Life-cycle theory of leadership. *Training and Development Journal*, 23, 26-34.

- Hersey, P., & Blanchard, K. (1982). *Management of organizational behavior* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Kerr, S., & Jermier, J. M. (1978). Substitutes for leadership: Their meaning and measurement. *Organizational Behavior and Human Performance*, 22, 375-403.
- Korman, A. K. (1966). Consideration, initiating structure, and organization criteria—A review. *Personnel Psychology*, 19, 349-361.
- Lewin, K. (1947). Frontiers in group dynamics: Concept, method, and reality in social science; social equilibria and social change. *Human Relations*, 1, 5-41.
- Liden, R. C., & Graen, G. (1980). Generalizability of the vertical dyad linkage model of leadership. *Academy of Management Journal*, 23, 451-465.
- Likert, R. (1967). *The human organization*. New York: McGraw-Hill.
- Maslow, A. (1954). *Motivation and personality*. New York: Harper & Row.
- McClelland, D. C. (1961). *The achieving society*. Princeton, NJ: Van Nostrand.
- McGregor, D. (1960). *The human side of enterprise*. New York: McGraw-Hill.
- Moore, L. I. (1976). The FMI: Dimensions of follower maturity. *Group and Organization Studies*, 1, 203-222.
- Reddin, W. J. (1967). The 3-D Management Style Theory. *Training and Development Journal*, 21, 8-17.
- Schein, E. H. (1970). *Organizational psychology* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.
- Schriesheim, C. A., & Kerr, S. (1974). Psychometric properties of the Ohio State leadership scales. *Psychological Bulletin*, 81, 756-765.
- Skinner, B. F. (1953). *Science and human behavior*. New York: MacMillan.
- Smith, P., Kendall, L., & Hulin, C. L. (1969). *The measurement of satisfaction in work and retirement*. Chicago: Rand-McNally.
- Stogdill, R., & Coons, A. (1957). *Leader behavior: Its description and measurement* (Research Monograph No. 88). Columbus: Ohio State University, Bureau of Business Research.
- Vecchio, R. P., & Gobdel, B. C. (1984). The vertical dyad linkage model of leadership: Problems and prospects. *Organizational Behavior and Human Performance*, 34, 5-20.
- Vroom, V. H., & Mann, F. C. (1960). Leader authoritarianism and employee attitudes. *Personnel Psychology*, 13, 125-140.

Received July 7, 1986

Revision received January 8, 1987

Accepted January 12, 1987 ■

Effects of Missing Application-Blank Information on Personnel Selection Decisions: Do Privacy Protection Strategies Bias the Outcome?

Dianna L. Stone

Department of Management, Bowling Green State University

Eugene F. Stone

Bowling Green State University

Using a $3 \times 2 \times 2$ experimental design and data from 188 managers and professionals, this study examined the main and interactive effects of information management strategy (missing information vs. no reported conviction vs. reported conviction), race of the applicant (White vs. Black), and job type (cashier vs. road laborer) on ratings of an applicant's qualifications and likelihood of job success. For the qualification criterion, there were significant main effects for information management strategy and job type. In the case of the success criterion, there were significant main effects for information management strategy and job type and two significant two-way interactions. Constraints on the generalizability of the findings to personnel decision making in actual organizational contexts are considered.

In response to growing concerns about the collection of irrelevant data by work organizations during preemployment screening activities (e.g., Privacy Protection Study Commission, 1977), a number of states have recently passed laws restricting the collection and use of arrest and conviction data for employment decisions (Smith, 1981, 1982). Rather than relying on legislative means for protecting privacy, some analysts (e.g., Harragan, 1983; Hayden & Novik, 1980) have suggested that job applicants simply not answer items on an application blank that they view as job-irrelevant or unfairly invasive of their privacy. However, as is explained below, (a) the use of this seemingly simple and effective strategy may serve to negatively bias judgments about an applicant's job suitability, (b) the bias may be more pronounced for Blacks and other minorities than it is for Whites, and (c) the degree of bias may be a function of the job for which the applicant has applied.

Although a considerable amount of research has focused on the improvement of personnel selection practices, this research has failed to consider the impact of personnel data-collection practices on either applicants or personnel decisions (see Guion, 1967; London & Bray, 1980; Messick, 1980; Schein, 1977). Therefore, the major purpose of this study was to assess the effects of applicants' information management strategies (providing or withholding requested information) on ratings of their job suitability. In addition, the study examined the extent to which this effect was moderated by applicant race and type of job for which the applicant applied.

As previously suggested, one strategy individuals can use to protect their privacy is to leave items unanswered on an application blank that they consider to be job-irrelevant or unfairly invasive. However, psychological theory and research (see Feldman, 1981; Rosch, 1977; Wyer & Carlston, 1979) suggests that the use of this strategy may have a negative or biasing effect on ratings of the applicant: Upon the discovery of missing information about arrests or criminal convictions, personnel decision makers may, on the basis of extant stereotypes, assume that an applicant refused to complete application items because he or she had something to hide (e.g., a previous conviction). As a result, they might rate the applicant in a manner that is comparable to an applicant who actually reports one or more convictions (cf. Feldman, 1981). On the basis of this theory, research, and reasoning, this study tests the hypothesis that an applicant who omits information about previous criminal convictions will be rated lower in terms of job suitability than an applicant who reports no record of criminal conviction (Hypothesis 1).

Empirical evidence (cf. Arvey, 1979b) suggests that applicant race may interact with other variables in determining judgments about job suitability. Thus, the effect of information management strategy on suitability ratings may be moderated by applicant race. More specifically, because the base rate of arrests and criminal conviction is higher for Blacks than Whites (Arvey, 1979a), theory and research on stereotyping (e.g., Ashmore & DelBoca, 1979; Bodenhausen & Wyer, 1985; Haefner, 1977; Hamilton, 1979; Schneider, Hastorf, & Ellworth, 1979) suggests that, compared with equally qualified White job applicants, Black applicants who fail to complete application-blank items about criminal conviction will receive lower suitability ratings. Thus, we test the hypothesis that applicant race and information management strategy will interact in determining applicant suitability ratings, so that the difference between the mean suitability ratings of Black and White applicants will be greater in instances when there is missing information about criminal conviction than it is when the applicants are conviction free (Hypothesis 2).

An earlier version of this article was presented at the meeting of the Southern Management Association, November 1985, in Orlando, Florida.

We thank Irwin L. Goldstein and two anonymous reviewers for helpful comments on a previous version of this article.

Correspondence concerning this article should be addressed to Dianna L. Stone, Department of Management, or Eugene F. Stone, Department of Psychology, Bowling Green State University, Bowling Green, Ohio 43403-0228.

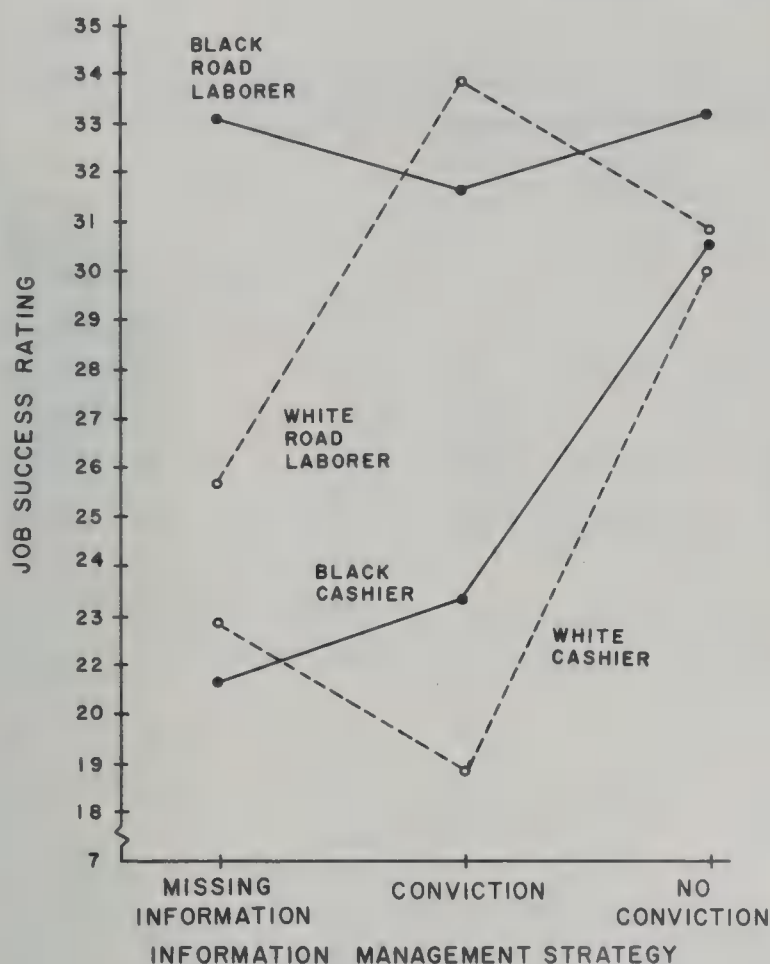


Figure 1. Effects of information management strategy, race, and job on ratings of job success.

As previously noted, in instances in which information about a criminal conviction is missing, a previous conviction may be assumed. If it is, the impact of missing information on suitability ratings may be moderated by the perceived relevance of a suspected conviction for job performance. In addition, perceived relevance may be a function of the general level of responsibility (for people, property, etc.) associated with a job. Consequently, we tested the hypothesis that information management strategy and job type will interact, so that ratings of applicant suitability will be influenced more negatively by missing information in the case of a high- than of a low-responsibility job (Hypothesis 3).

Method

Subjects

Subjects were 188 (119 men and 69 women) managers or professionals recruited from executive programs and MBA classes at two large Northeastern universities. Their average age was 33.89 years. Sixty-six (35.1%) of the subjects were personnel managers. One hundred four (55.3%) reported that their work involved personnel selection.

Design and Procedures

Subjects were randomly assigned to 1 of 12 experimental conditions corresponding to unique manipulations of information management

strategy, applicant race, and job type. In a single session, subjects were (a) given and asked to complete an informed consent agreement, (b) provided a packet of materials containing a job description, an application blank completed by a fictitious, male job candidate, and a rating form, (c) asked to review sequentially the job description, the application blank, and complete the rating form, (d) asked to complete a questionnaire containing manipulation check items, and (e) debriefed.

Manipulations

Information management strategy. Information management strategy was varied by altering the applicant's responses to the application-blank question that read "Have you ever been convicted of a crime?" These alterations included (a) a "no conviction" response, (b) a "yes" response, accompanied by the applicant's notation that once he had been convicted of petty theft and received a suspended sentence, and (c) a missing response.

Race. In order to manipulate race, the application-blank question dealing with this variable indicated that the job applicant was either White or Black.

Job type. Job type (and corresponding responsibility level) was manipulated by varying the title and description of the job for which the individual applied. In the road-laborer (low-responsibility) condition, the job duties consisted of spreading tar and gravel on roadbeds under the direct supervision of a foreman. In the cashier (high-responsibility) condition, the job entailed handling and accounting for customer checks and cash payments.

Measures

Two aspects of suitability, qualification for the job (*qualification*) and potential job success (*success*), were indexed with a questionnaire having 7-point (*strongly disagree* to *strongly agree*) response possibilities. Qualification was indexed with six items (e.g., "I feel that this applicant is very well suited for the job") ($\alpha = .82$). Success was measured with seven items (e.g., "The chances of this person being successful on the job are quite high") ($\alpha = .87$).

Analyses

Hypotheses about the suitability of the applicant were tested separately for the indices of qualification and success using hierarchical multiple regression (cf. Stone & Hollenbeck, 1984). In these analyses, scores on the dependent variable were regressed on dummy- or effect-coded variables representing information management strategy (1 = *no conviction*, 0 = *conviction*, and -1 = *missing*), race of the applicant (1 = *White* and -1 = *Black*), job type (1 = *road laborer* and -1 = *cashier*), and crossproduct terms carrying information about interactions between these independent variables. In instances in which an initial analysis showed no interaction effects, a second analysis was performed that estimated only the main effects of the independent variables. Type I error rates of .05 and .10, respectively, were used to test the significance of main and interactive effects.

In order to illustrate the bases for the significant effects found in the regression analyses, *t*-tests with a .05 Type I error rate were used to contrast the means of specific treatment conditions.

Results

Table 1 presents descriptive data on the suitability measures for selected experimental conditions.¹ Figure 1 shows a plot of

¹ A table showing individual cell means is available from the authors.

Table 1
Job Suitability Ratings for Selected Treatment Conditions

Treatment condition/measure	Information management strategy								
	Missing information			Conviction			No conviction		
	<i>M</i>	<i>S</i> ²	<i>N</i>	<i>M</i>	<i>S</i> ²	<i>N</i>	<i>M</i>	<i>S</i> ²	<i>N</i>
Black									
Success	27.65	88.17	31	27.61	63.68	31	32.28	75.00	29
	(a)			(b)			(c)		
Qualification	24.48	70.06	31	23.97	58.83	31	26.59	80.28	29
	(d)			(e)			(f)		
White									
Success	24.39	79.21	36	26.83	88.55	30	30.53	97.42	30
	(g)			(h)			(i)		
Qualification	22.50	66.42	36	23.00	105.88	30	5.27	49.14	30
	(j)			(k)			(l)		
Cashier									
Success	22.38	74.30	32	21.57	62.41	30	30.90	63.20	29
	(m)			(n)			(o)		
Qualification	19.00	43.03	32	16.40	39.69	30	22.14	63.20	29
	(p)			(q)			(r)		
Road laborer									
Success	29.11	74.30	35	32.71	26.11	31	31.87	109.83	30
	(s)			(t)			(u)		
Qualification	27.46	57.76	35	30.35	24.40	31	29.57	38.44	30
	(v)			(w)			(x)		
Across job and race conditions									
Success	25.90	84.64	67	27.10	74.61	62	31.39	85.75	59
	(y)			(z)			(aa)		
Qualification	23.42	68.06	67	23.48	79.53	62	25.92	63.68	59
	(bb)			(cc)			(dd)		

Note. *S*² is the variance for sample. Letters in parentheses identify the values used in the *t* tests reported in Table 3.

the twelve cell means for the success variable. (Because the plot for the qualification variable was similar in appearance it is not presented here.) Table 2 shows results of the multiple regression/correlation analyses for the qualification and success variables. Table 3 presents the results of *t* tests comparing specific treatment conditions to one another.²

Consistent with Hypothesis 1, regression analyses (see Table 2) showed the information management strategy manipulation had significant effects on both the qualification and the success criteria. Mean scores for the same criteria were greater in the no-conviction than in the missing-information condition (see Table 3). Moreover, ratings on success and qualification for the missing-information and the reported-conviction conditions did not differ (see Tables 1 and 3).

Results of the regression analyses designed to test Hypothesis 2 showed a significant Race × Information Management Strategy interaction for the success criterion, but not for the qualification measure. However, contrary to our prediction, the results in Tables 1 and 3 demonstrate that, for the missing-information condition, success ratings were higher for the Black than for the White applicant.

In agreement with Hypothesis 3, the regression analysis for the success variable showed a significant Job × Information Management Strategy interaction effect (see Table 2). Moreover, results of *t* tests showed that the difference between means

on the success criterion (no conviction vs. missing information) was greater for the cashier than for the road-laborer condition. (see Tables 1 and 3). Note that in the case of the qualification criterion, the regression analysis failed to reveal a significant Job × Information Management Strategy interaction.

Apart from the findings already considered, the results in the tables reveal several other findings of interest. First, job and race interactively influence ratings of success. The nature of this interaction is suggested by the mean success ratings displayed in Figure 1. They reveal that, although Blacks were uniformly rated as more likely to succeed than Whites, the difference between means for these race groups was greatest in the case of the road-laborer job. Second, there was a main effect of job type on both success and qualification ratings. Consistent with intuition, the lower the general cognitive and social demands of a job, the more suitable any given applicant is perceived to be. Third, there is a main effect of race on both measured-suitability indices. Contrary to what is suggested by extant racial stereotypes, however, Blacks are generally rated as more suitable for the studied jobs than are Whites.

² A table showing degrees of freedom for these tests is available from the authors.

Table 2
Regression of Suitability Ratings on Information Management Strategy, Applicant Race, and Job Type

Independent variable	Dependent variable			
	Success ^a		Qualification ^b	
	β	<i>t</i>	β	<i>t</i>
Information management strategy (I)	.23	3.47** ^c	.12	2.11** ^c
Job type (J)	.62	4.62** ^c	.59	10.14** ^c
Race of applicant (R)	-.10	1.53 ^d	-.11	1.80 ^d
J \times R	-.34	2.38** ^d	—	—
J \times I	-.11	1.66* ^d	—	—
R \times I	—	—	—	—

Note. A dash (—) signifies that a variable was not included in the final regression analysis.

^a $R = .477$, $F(5, 180) = 10.62$, $p < .05$. ^b $R = .610$, $F(3, 184) = 36.35$, $p < .05$. ^c One-tailed test of significance used. ^d Two-tailed test of significance used.

* $p < .10$. ** $p < .05$.

Discussion

Results of the study showed that applicants who fail to respond to application-blank items dealing with criminal conviction are viewed as less suitable for jobs than those who report no conviction. Thus, individuals who attempt to protect their privacy through a nonresponse strategy will suffer costs that seemingly were not considered by various advocates of nonlegislative privacy-protection mechanisms (e.g., Harragan, 1983; Hayden & Novik, 1980). If, as this study's results suggest, a potential employer views a nonresponse to an application-blank item as an attempt to conceal facts that would reflect poorly on an applicant, then applicants (especially those free of previous arrests, convictions, or other stigmatizing attributes) would be ill advised to not respond. Put simply, refusing to respond to items does not appear to be a viable strategy for applicants in protecting their actual or desired rights to privacy. Consequently, if legislators believe that the merits of protecting individuals' rights to privacy outweigh the perceived needs of employers to know sensitive information (e.g., about arrests or criminal convictions), then there may be merit in adopting one or more of several previously suggested privacy protection strategies (cf. Privacy Protection Study Commission, 1977; D. Stone, 1981; E. Stone, 1980; Stone, Gueutal, Gardner, & McClure, 1983). One possibility is the passage of legislation to curb the collection of information that is sensitive, privacy invasive, and not job-relevant.

Although we hypothesized that race and information management strategy would influence ratings of job suitability interactively, the analyses failed to show evidence of such an effect. Two interpretations of these findings appear plausible, both of which are based on the fact that all subjects were employed by racially heterogeneous organizations in urban, northeastern cities. One is that subjects did not base their judgments about applicant suitability on racial biases. A second is that even though subjects may indeed have had racial biases, they chose not to

reveal them to researchers. It is unclear, therefore, that research employing subjects from other regions of the United States would fail to find Race \times Information Management Strategy interactions or that actual personnel decisions are unaffected by racial biases.

Although not hypothesized, we found that job and race determined success ratings interactively. More specifically, although the rated success of Blacks and Whites did not differ for the cashier job, Blacks were rated above Whites for the road-laborer position. This finding may be based on the stereotypical belief that Blacks are more suited for physically demanding work than are Whites. If this is indeed the case, then other beliefs connected with racial stereotypes may work to the disadvantage of Blacks when they are being considered for other types of positions (e.g., professional or managerial jobs). For such positions, race-related biases may operate to the disadvantage of Black (and other minority) applicants. Research that deals with this issue is needed.

Table 3
Differences Between Suitability Ratings for Selected Treatment Conditions

Treatment conditions compared	Suitability index	Means compared	<i>t</i>	ω^2
Findings relevant to Hypothesis 1				
NC vs. MI	S	aa vs. y	3.33**	.099
	Q	dd vs. bb	1.73*	.016
MI vs. CO	S	y vs. z	0.76	.000
	Q	bb vs. cc	0.04	.000
Findings relevant to Hypothesis 2				
NC-BA vs. NC-WA	S	c vs. i	0.72	.000
	Q	f vs. l	0.63	.000
MI-BA vs. MI-WA	S	a vs. g	-1.45	.016
	Q	d vs. j	-0.98	.000
CO-BA vs. CO-WA	S	b vs. h	-0.35	.000
	Q	e vs. k	-0.42	.000
Findings relevant to Hypothesis 3				
NC-CA vs. MI-CA	S	o vs. m	4.02**	.199
	Q	r vs. p	1.67*	.028
NC-RL vs. MI-RL	S	u vs. s	1.15	.005
	Q	x vs. v	1.23	.008
NC-RL vs. NC-CA	S	u vs. o	0.40	.000
	Q	x vs. r	3.99**	.202
MI-RL vs. MI-CA	S	s vs. m	3.19**	.120
	Q	v vs. p	4.89**	.255
CO-RL vs. CO-CA	S	t vs. n	6.52**	.405
	Q	w vs. q	9.60**	.599

Note. BA = Black applicant; CA = cashier job; CO = conviction; MI = missing information; NC = no conviction; RL = road laborer job; S = success criterion; Q = qualification criterion; WA = White applicant. *d*/*f*s for *t* tests were based on formula 10.17.1 in Hays (1973). This formula estimates *d*/*f*s in cases in which a *t* test is based on groups having unequal sample sizes and unequal variances. Omega-square estimates are based on formula 10.20.1 of Hays (1973). Letters in the Means-compared column refer to the associated means that are reported in Table 1.

* $p < .05$. ** $p < .005$.

Although our sample consisted of working subjects, many of whom made personnel decisions, these subjects rated the suitability of simulated as opposed to real job applicants. Consequently, research is needed to assess the generalizability of our findings. In addition, research is needed that examines how missing information on application blanks affects actual personnel decisions. This research should consider the nature of information that is omitted (e.g., arrest, medical, sex, national origin, drug use), the nature of the job for which an individual is applying (e.g., managerial vs. nonmanagerial, blue vs. white collar), and the job relevance of the missing information.

References

- Arvey, R. D. (1979a). *Fairness in selecting employees*. Reading, MA: Addison-Wesley.
- Arvey, R. D. (1979b). Unfair discrimination in the employment interview: Legal and psychological aspects. *Psychological Bulletin*, 86, 736-765.
- Ashmore, R. D., & DelBoca, F. K. (1979). Sex stereotypes and implicit personality theory: Toward a cognitive-social psychological conception. *Sex Roles*, 5, 219-248.
- Bodenhausen, G. V., & Wyer, R. S., Jr. (1985). Effects of stereotypes on decision-making and information-processing strategies. *Journal of Personality and Social Psychology*, 48, 267-282.
- Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology*, 66, 127-148.
- Guion, R. M. (1967). Personnel selection. *Annual Review of Psychology*, 18, 208-209.
- Haefner, J. E. (1977). Race, age, sex, and competence as factors in employer selection of the disadvantaged. *Journal of Applied Psychology*, 62, 199-202.
- Hamilton, D. L. (1979). A cognitive-attributational analysis of stereotyping. In L. Berkowitz (Ed.) *Advances in experimental social psychology* (Vol. 12, pp. 53-84). New York: Academic Press.
- Harragan, B. L. (1983). Getting ahead: Career priorities. *Working Woman*, 8, 38.
- Hayden, T., & Novik, J. (1980). *Your rights to privacy*. New York: Avon Books.
- Hays, W. L. (1973). *Statistics for the social sciences* (2nd ed.). New York: Holt, Rinehart & Winston.
- London, M., & Bray, D. W. (1980). Ethical issues in testing and evaluation of personnel decisions. *American Psychologist*, 35, 890-901.
- Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, 35, 1012-1027.
- Privacy Protection Study Commission. (1977). *Personal privacy in an information society*. Washington, DC: U.S. Government Printing Office.
- Rosch, E. (1977). Human categorization. In N. Warren (Ed.), *Studies in cross-cultural psychology* (Vol. 1, pp. 3-49). New York: Academic Press.
- Schein, V. E. (1977). Individual privacy and personnel psychology: The need for a broader perspective. *Journal of Social Issues*, 33, 154-168.
- Schneider, D. J., Hastorf, A. H., & Ellsworth, P. C. (1979). *Person Perception*. Reading, MA: Addison-Wesley.
- Smith, R. C. (1981). *Compilation of state and federal privacy laws*. Washington, DC: Privacy Journal.
- Smith, R. C. (1982). *Supplement to state and federal privacy laws*. Washington, DC: Privacy Journal.
- Stone, D. L. (1981). The effects of the valence of outcomes for providing data and the perceived relevance of the data requested on privacy-related behaviors, beliefs, and attitudes. *Dissertation Abstracts International*, 42, 3731-A. (University Microfilms No. 82-00, 737).
- Stone, E. F. (1980). *Testimony presented at U.S. Labor Department Hearings on Workplace Privacy* (Working Paper No. 7). West Lafayette, IN: Purdue University, Information Privacy Research Center.
- Stone, E. F., Gueutal, H. G., Gardner, D. G., & McClure, S. (1983). A field experiment comparing information-privacy values, beliefs, and attitudes across several types of organizations. *Journal of Applied Psychology*, 68, 459-468.
- Stone, E. F., & Hollenbeck, J. R. (1984). Some issues associated with the use of moderated regression. *Organizational Behavior and Human Performance*, 34, 195-213.
- Wyer, R. S., & Carlston, D. E. (1979). *Social cognition, inference and attribution*. Hillsdale, NJ: Erlbaum.

Received August 6, 1985

Revision received January 29, 1987

Accepted January 31, 1987 ■

Stability of Skilled Performance Across Time: Some Generalizations and Limitations on Utilities

Rebecca A. Henry and Charles L. Hulin
University of Illinois at Urbana-Champaign

We review two closely related phenomena in the study of ability and performance: (a) decreasing predictive validity in ability-performance relations when studied longitudinally and (b) the superdiagonal stability matrices (i.e., simplex-like matrices) found in the correlations between repeated trials on a variety of tasks. We discuss briefly two models that explain these phenomena, a changing-task model and a changing-subject model, and report an empirical study in which (a) the generality of the simplex phenomenon in an area of human performance not previously studied and (b) whether the decrease in correlations across time would eventually stabilize (a hypothesis of the changing-task model) were investigated. Intercorrelation matrices of major league baseball performance across 10 years (four measures) indicate that the correlations systematically decrease over time, with no evidence of stabilizing. We discuss the implications of these empirical findings and conceptual developments for utilities of selection programs.

A ubiquitous finding in the study of relations of ability and performance is the decreasing predictive validity that occurs as a function of time and interpolated practice. Correlations between ability measures, assessed at Time i , and performance, assessed successively at Times $i + 1, i + 2, \dots, i + k$, decrease regularly as a function of the ordinal position of the performance assessment (Adams, 1957; Alvares & Hulin, 1972; Dunham, 1974; Fleishman, 1960; Fleishman & Hempel, 1954, 1955; Fleishman & Parker, 1959; Fleishman & Rich, 1963; Humphreys, 1968; Lin & Humphreys, 1977). Studies done longitudinally with constant samples of subjects, almost without exception, display this general trend.

A second phenomenon, argued to be closely related to the first, is the superdiagonal stability matrix that exists in the intertrial correlations of learning tasks or in time-period by time-period correlations of performance on a variety of tasks (Dennis, 1954, 1956; Humphreys & Davey, 1984; Humphreys, Davey, & Park, 1984; Humphreys & Parsons, 1979; Wilson, 1983). This superdiagonal pattern, characterized by the highest correlations occurring between adjacent trials with successive correlations decreasing monotonically as a function of the number of intervening trials or time periods, was recognized and discussed more than 50 years ago by Perl (1934) in the context of motor-skill acquisition. Satisfactory, generally accepted explanations for these two phenomena have yet to be developed.

In a review of the literature, Alvares and Hulin (1972) suggested that the occurrence of the decreasing validity coefficients and superdiagonal stability matrices are manifestations of the

same process. To arrive at this conclusion, they assumed that early performance on a task and performance on an ability test are manifestations of the same underlying phenomenon. Both assessments reflect individuals' current repertoires of skills, knowledge, and abilities and are consistent with definitions of human ability (Hulin & Humphreys, 1980; Humphreys, 1985) that stress these as the basic components of abilities, including intelligence. Given this assumption, the first row of the time-period by time-period stability matrix is analogous to a vector of validity coefficients predicting performance across time and practice. If job samples are used as predictor measures, the first row of a stability matrix is identical to a vector of predictive validities.

Humphreys (1960) was the first to apply the term *simplex*, borrowed from Guttman (1955), to matrices of longitudinal data displaying this pattern. Guttman used the term to describe an intercorrelation matrix of items measuring the same factor, but differing in difficulty. If items are arranged from least to most difficult, as in a Guttman scale, the resulting correlation matrix will display a characteristic pattern of correlations. Adjacent correlations will be high, and the correlations will decrease monotonically as a function of the ordinal distance between the items. In a true simplex, $\rho_{ik,j} = 0$, for all i, j , and k , $i < j < k$.

Generality of the Phenomenon

The two major areas of study in which these phenomena have been found are (a) prediction of performance and (b) growth and development. In prediction studies a wide range of psychomotor abilities have been investigated as predictors of discriminant reaction time (Adams, 1957; Fleishman & Hempel, 1954, 1955), tracking (Dunham, 1974; Fleishman & Parker, 1959), rotary pursuit (Fleishman, 1960), two-hand coordination (Fleishman & Rich, 1963), and student pilot performance (Alvares & Hulin, 1973). Predictions of academic performance in college (Humphreys, 1968; Humphreys & Taber, 1973) and

We would like to thank Lloyd Humphreys, Fritz Drasgow, and Mary Roznowski for reading and commenting on an earlier draft of this article. Their efforts contributed significantly to the quality of the final product.

Correspondence concerning this article should be addressed to Rebecca A. Henry, Department of Psychology, University of Illinois, 603 East Daniel Street, Champaign, Illinois 61820.

graduate or professional school (Carlson & Werts, 1976; Lin & Humphreys, 1977; Powers, 1982) have also shown this characteristic pattern of decreasing validities when standard aptitude tests, previous academic performance, or both combined are used as predictors. Prediction studies that have spanned decades have also displayed the superdiagonal pattern: salary as a predictor of salary over 20 years in a sample of engineers in the same company (Brenner & Lockwood, 1965) and productivity of scientists across four decades (Dennis, 1954, 1956).

Growth and development studies, primarily in the area of human intelligence, have also found the superdiagonal, simplex-like pattern. Anderson (1939, 1940) was the first to observe it. Recently, three studies have shown good fits of true simplex matrices to general intelligence test data after correcting for attenuation (Humphreys & Davey, 1984; Humphreys, Davey, & Park, 1984; Humphreys & Parsons, 1979).

The ubiquity of these two phenomena and what they do and do not imply about human ability and performance needs to be stressed. Both the decreasing validity coefficients and the characteristic superdiagonal stability matrices have been observed in nearly every longitudinal study involving repeated assessments of performance over time. The lower asymptote of the validity curve appears to be approximately zero, the decrement appears to be determined by the size of the initial validity, and the decrement in validity appears to be greater if performance is assessed objectively rather than by means of supervisor or peer ratings. There have been no published studies of trial-by-trial or time-period by time-period stability matrices that fail to show the simplex-like structure in the intercorrelations. Both of these sets of empirical findings suggest dynamic growth and change models rather than assumptions of invariant or fixed human abilities and skilled performance. They also imply that this growth is not simple, but rather that systematic changes in the rank order of individuals in terms of ability or performance are to be expected.

It is also important to note what the data do not imply. There are no implications about relations between ability and performance when they are studied cross sectionally. Indeed, differences in human ability assessed at Time i should be related to performance at Time i even if different cross-sectional samples are studied at different levels of experience. The published data are relevant to relations between ability at Time 1 and performance or ability at Times 2, 3, . . . , i .

Explanations for Decreasing Validity Coefficients and Performance Stability Matrices

Alvares and Hulin (1972) labeled two alternative models of changes in ability–performance relations as changing-task and changing-subject models. The changing-task model (Fleishman & Hempel, 1954) states that the repertoire of abilities needed to perform a task (i.e., the task structure) changes with time or interpolated practice; the amounts of these abilities possessed by individuals remain constant. Conversely, the changing-subject model (Adams, 1957; Alvares & Hulin, 1972; Corballis, 1965) asserts that it is the individual's level of ability that changes over time, whereas the specific abilities required for performance on the task remain constant across time.

Both of these models can be described in terms of the compo-

nents of individual performance. This may be expressed as follows:

$$Y_{ij} = \sum (a_{jk}x_{ikj}) + e_{ij} + s_j, \quad (1)$$

where Y_{ij} = performance score of the i th person on j th trial; a_{jk} = weight of k th common ability for determining performance on the j th trial; x_{ikj} = amount of k th common ability possessed by i th person at time of j th trial; e_{ij} = error of measurement for i th person on j th trial; s_j = a factor specific to trial j ; and $\text{Cov}(e_{ij}e_{ik}) = 0$, $\text{Cov}(e_{ij}x_{ikj}) = 0$, and $E(e_{ij}) = 0$.

It is evident that either a change in a_{jk} or a change in x_{ikj} could account for the change in an individual's performance score across trials. The changing-task model asserts that a_{jk} changes, whereas the changing-subject model asserts that x_{ikj} changes; thus both models can account for the decreasing validity coefficients and change in performance levels across time.

The research and analytic methods used to support the changing-task model have been criticized on a number of points. Adams (1957) criticized the methods used in the empirical research on which the changing-task interpretations were based because factor analysis ignores the conceptual and experimental distinctions between predictors and criteria. Factors defined by predictor variables were allowed to influence the definitions of factors defined by performance measures; performance and predictor variable spaces were not differentiated in the factor analytic procedures. Bechtoldt (1960, 1961) noted that neither strengths of relations nor changes in strengths of relations across trials can be assessed rigorously or studied using these procedures. The inability of the analytic methods used to provide any rigorous evidence about the degree or amount of change in a relation between sets of variables seems unfortunate in light of the importance these changes assumed in the conceptual development of the changing-task model.

Humphreys (1960) observed that when a minimum rank factor analytic technique is applied to a simplex-like matrix, the result will always be at least a two-factor solution, with one factor defined by early trials and the other by later trials. This pattern, which was used to develop the changing-task model, will occur even when only one common factor has generated the observed data (as is the case with a perfect simplex).

Empirical tests of the changing-subject model are less frequent. Adams (1957) used multiple regression techniques to predict discriminant reaction time from simple reaction time and the Airman Classification Battery. From the initial to final trials he found a slightly smaller multiple correlation, as well as systematic changes in regression weights for the predictor variables. He accounted for these changes in terms of the changing-subject model. However, he did not report whether there were any mean changes or changes in the rank ordering of subjects on ability measures as a function of performance on the criterion task. Thus, a changing-task model could also account for the decreasing correlations and changes in regression weights.

Alvares and Hulin (1973) found in posttraining comparisons of ability measures obtained from a pilot trainee group and a control group that did not receive flight training, that individual ability levels on nine tests related to flying proficiency showed change due to the flight experience and training. Scores on three tests of tool and mechanical knowledge unrelated to flying proficiency did not change. This study demonstrated differential

changes in human abilities as a function of the relation of the ability measures to flight proficiency and as a function of practice on the criterion task.

Dunham (1974) tested whether either explanation alone or a combination of the two should be used to account for the change in predictability of performance over time. He used visual acuity and simple reaction time as predictors because visual acuity should not change as a function of an hour's practice on a tracking task; changes in simple reaction time might be expected as a function of practice on a discriminant reaction time test. Support was found for the changing-subject model, although a significant portion of the change in predictability could not be accounted for with this model.

In summary, the empirical evidence required by the changing-task model has not been conclusive. Limited support has been found for a changing-subject effect, but this model cannot account for all of the observed decrease in validity across time or practice (Alvares & Hulin, 1972; Dunham, 1974).

Performance of Professional Baseball Players

To investigate the generality of the simplex phenomenon in an area of human performance not previously studied, archival data were collected on professional baseball players over a 10-year period. This sample was chosen because professional baseball players are highly selected, highly motivated individuals performing well-learned tasks. The tasks of professional baseball players seem appropriate to test one hypothesis derived from the changing-task model.

One assumption made by advocates (Fleishman & Hempel, 1954, 1955; Fleishman & Parker, 1959) of the changing-task model is that the systematic changes in task requirements will stabilize when the tasks become overlearned. The tasks of professional baseball players can be argued to be overlearned by the time the players have worked their way through little leagues; high school; occasionally, college ball; summer instructional leagues; minor leagues; and have finally reached the major leagues. An additional desirable feature of such a sample is that directly competitive tasks have not been studied in past research in this area.

Two measures of offensive performance based on yearly averages or totals were used: batting average (hits divided by official at bats), and total runs produced (runs scored, plus runs batted in, minus home runs). Neither measure is perfect. Runs produced is probably a better assessment of the total offensive production of a player than is simple batting average. It is, however, influenced by the opportunities one has to drive in runs; batters in the eighth and ninth places in the batting order have fewer opportunities than those in the second through fifth spots in the order. However, our emphasis is on stability and change over time, not on absolute performance. These opportunity factors, unless they change systematically, should increase the apparent stability of performance because established lead off batters, clean up batters, and those at the lower end of the batting order tend to hit in the same positions from one year to the next. The similarity of the patterns in the batting average and total runs produced matrices suggests that these opportunity factors do not confound our results.

Defensive measures were collected only for pitchers. These

included earned run average (earned runs allowed, multiplied by nine, divided by innings pitched) and a composite measure of pitcher performance on the basis of earned runs (ER), strike outs (SO), walks (W), and innings pitched (IP):

$$\frac{[4(ER) + W - SO]}{IP}$$

All of the players who satisfied the following criteria were included in the samples: (a) 10 years of continuous play, with the 1st year being their last year in the minor leagues and Years 2 through 10 being in the major leagues; (b) for offensive performance, a minimum of 100 annual at bats ($n = 94$); and (c) for pitchers, at least 50 innings pitched ($n = 38$). All of these measures were taken from *The Sporting News Official Baseball Register* (Siegel, 1985), a cornucopian source of performance measures.

Unstable and marginal players (those who returned for occasional encore performances in the minor leagues) were excluded because of statistical problems associated with comparing performance measures across major and minor leagues. The sample sizes (94 and 38) are small, but are large enough to produce stable results. It is our speculation that the superdiagonal stability matrices would have been even more pronounced had we been able to include marginal players, and had our selection criteria allowed larger sample sizes. In this case, sampling fluctuations around the expected correlations would have been less pronounced. Including marginal players as well as established journeymen, stars, and superstars would have increased the spread of talent and the variance in the samples. It would have also weakened the argument that these tasks are overlearned in our samples.

Correlation matrices were obtained for these four measures (two offensive and two defensive). An additional matrix was calculated for pitching performance by averaging the correlations (after converting to Fisher z scores) for the three separate components of the composite measure (strike outs, walks, and earned runs). With few exceptions, superdiagonal stability matrices for all of the performance measures were found. The total runs-produced matrix and the composite pitcher-performance matrix are shown in Tables 1 and 2.

Both matrices have the characteristic superdiagonal stability form. The highest correlations occur between pairs of adjacent years. Averages of these adjacent-year correlations are $+ .51$ and $+ .67$ for total runs produced and pitcher performance, respectively. With a few exceptions, these correlations decrease regularly as a function of the number of intervening years. The predictive validity of minor league performance decreases substantially through the ninth year in the majors, especially in the sample of pitchers. In addition, there is little tendency for the decrease in correlations to stabilize during the later years in the majors.

Current Policy and Policy Implications

The studies in the empirical literature that were summarized briefly, and our baseball player performance analysis, illustrate the generality of the simplex phenomenon. Occurrences of these matrices extend beyond perceptual-motor skills and the traditional organizational and academic performance assess-

Table 1
Intercorrelations of Annual Total Runs Produced

Year	1	2	3	4	5	6	7	8	9	10
1. Minor (last)	—	.31	.20	.14	.18	.08	.12	.15	.22	.12
2. Major 1		—	.53	.20	.19	.27	.18	.08	.22	.11
3. Major 2			—	.50	.44	.42	.32	.24	.19	.14
4. Major 3				—	.55	.51	.35	.35	.35	.19
5. Major 4					—	.59	.41	.28	.32	.33
6. Major 5						—	.47	.34	.46	.27
7. Major 6							—	.52	.39	.40
8. Major 7								—	.56	.49
9. Major 8									—	.50
10. Major 9										—

Note. *N* = 94. *r* > .17, *p* < .05.

ments. Indeed, they have been found in every area of performance studied. Laboratory studies lasting 1 hour, and longitudinal studies ranging from 1 or 2 weeks to 40 years, have consistently found the characteristic pattern of results. Unfortunately, the pervasiveness of the phenomenon has stimulated neither rigorous empirical and theoretical research aimed at explaining the causes nor even a widespread recognition of the temporal limitations of selection programs. Limited support has been found for the changing-subject explanation, although this model cannot account for all of the decreases in predictive validities. These findings stand in contrast to selection policies in education and industry that assume that abilities and performance are stable across time.

Acknowledgment of the superdiagonal matrices and the transitory nature of predictive validities across all areas studied (in terms of both over-learned skills studied here and newly learned skills; e.g., Fleishman, 1960) would complicate the tasks researchers must undertake and the questions they must address. Utilities of selection programs must consider the time-bound nature of the estimates. Predictive validity models of performance or skill acquisition that take the apparent temporal factors into consideration would seem to be an inevitable result of ■ recognition of these consistent findings. These models have not yet been developed.

In spite of these findings, most trait-theoretic models of human ability and performance do not normally include provi-

sions for lawful change as part of the models. Selection policies in industry and education reflect this assumption. Indeed, Hunter & Hunter (1984) have argued that ability is the best predictor of job performance because other measures such as “bio-data keys often suffer decay in validity over time” (p. 87). Bio-data keys do indeed suffer from decay in validity over time. The implicit and erroneous assumption by Hunter and Hunter seems to be that ability measures do *not* suffer from such a decay. In this same context, Hunter and Hunter computed the utility of a testing program as

$$U = NT r_{xy} \sigma_y M, \tag{2}$$

where *U* = the utility of a selection program; *N* = the number of applicants hired; *T* = the expected tenure of those hired; *r_{xy}* = test validity; *σ_y* = standard deviation of evaluated performance scores (evaluated on a dollar scale); and *M* = mean ability of those hired.

Hunter and Hunter (1984) noted, “even the smallest drop in validity would mean substantial losses (in utility)” (p. 89). The use of only one “validity” evaluated during the initial time periods of performance assumes that validity does not decrease systematically. It therefore produces an overestimate of the utility of the selection program. Other examples of these assumptions to estimate utility may be found in Hunter and Schmidt (1982), Schmidt, Hunter, McKenzie, and Muldrow (1979), and in ear-

Table 2
Intercorrelations of Overall Pitcher Performance

Year	1	2	3	4	5	6	7	8	9	10
1. Minor (last)	—	.49	.24	.19	.09	.12	.06	.03	.14	.08
2. Major 1		—	.51	.42	.36	.32	.27	.25	.45	.23
3. Major 2			—	.71	.49	.43	.26	.37	.41	.31
4. Major 3				—	.65	.56	.43	.57	.53	.48
5. Major 4					—	.74	.61	.63	.68	.57
6. Major 5						—	.71	.70	.70	.61
7. Major 6							—	.72	.71	.69
8. Major 7								—	.76	.79
9. Major 8									—	.70
10. Major 9										—

Note. *N* = 38; *r* > .27, *p* < .05.

lier treatments of utility (Brogden & Taylor, 1950; Cronbach & Gleser, 1965; Taylor & Russell, 1939). The use of this form of the utility equation cannot be justified by the empirical data. This distortion of the empirical data inflates estimates of the usefulness of selection tests.

Many examples of selection and placement programs based on early ability and performance measures are available. The selection of pilots by all three services in the U.S. Department of Defense is based substantially on a mixture of measures taken from traditional ability tests, as well as on measures taken from *early* performance on perceptual and psycho-motor tasks.

Aptitude tests place young children into normal and special education classes (PASE v. Hannon, 1980; Larry P. v. Riles, 1979). Children in these classes are placed there more or less permanently; their performance is not reviewed periodically, as the empirical evidence suggests it should be, to determine if they have improved sufficiently to be moved back into regular classrooms. Indeed, it was this permanent placement aspect of the special education classes, rather than test bias per se, that was responsible for the federal courts' decision in Washington, DC (Hobson v. Hansen, 1967; Smuck v. Hobson, 1969). Such permanent placement decisions seem to reflect a belief in the fixed nature of traits. Paradoxically, it does not suggest a belief in the efficiency of the special education programs in schools. It also ignores the implications of the empirical data in this area. The validities of the Scholastic Aptitude Test for predicting success as an undergraduate often approach zero by the fifth or sixth semester (Humphreys, 1985).

Scores on the Law School Admission Test (LSAT), Graduate Management Admission Test (GMAT), Medical College Admission Test (MCAT), and Graduate Record Examination (GRE) are used to predict success in graduate or professional schools despite the often trivial long-term predictive validities. Predictions are not updated on the basis of performance during definable phases of postbaccalaureate training; decisions about future phases of training are not made on the basis of these updated predictions.

Professors are appointed to tenured rank positions on the basis of the quantity and quality of publications during their first 4 to 6 years of professional work. Professional athletes are chosen and signed to lucrative long-term contracts on the basis of their performance in college or in minor leagues.

These decisions, and myriad others, are made on the assumption that long-term performance can be predicted from an aptitude test, a motor skills battery, or early performance during the initial years on a job or in a training situation.

Instability and change in nearly all areas of human performance, skills, and measures of general ability are more to be expected than is stability. A changing-subject model explaining such change in terms of lawful changes in human abilities was discussed. Research on the basic validity of this model of change is scarce. What evidence is available suggests that the model is a valid but incomplete explanation for the observed decreases in predictive validities and performance changes. The generality of the phenomena and the implications of the findings for practice in education and organizations seem to point to the need for a serious examination of some of the widespread assumptions we make about human traits and skilled performance levels.

We do not argue that psychological scales and ability measures have trivial or meaningless validities for predicting job, academic, or other skilled performance. Indeed, such measures often achieve impressive validities. When used in conjunction with other measures (e.g., grade point average, bio-data), the usefulness of the selection program is difficult to question. We intended to raise questions about long-term predictions of performance. The failures of researchers to develop models that address long-term predictions and build into predictive equations measures that will reflect expected changes in the abilities of the selected employees or students is a source of serious concern. Assumptions about constant validities across time to estimate the utility of selection programs are counter to 50 years of evidence. A changing-subject model that will account for much of these data has been developed, and empirical data verifying some predictions made on the basis of this model have been supportive. Testing has been shown to be of significant benefit to organizations. However, improvements in test utilities can best be accomplished through a recognition of the dynamics of human performance and ability rather than assumptions about constant abilities and validities.

References

- Adams, J. A. (1957). The relationship between certain measures of ability and the acquisition of a psychomotor response. *Journal of General Psychology*, 56, 121-134.
- Alvares, K. M., & Hulin, C. L. (1972). Two explanations of temporal changes in ability-skill relationships: A literature review and theoretical analysis. *Human Factors*, 14, 295-308.
- Alvares, K. M., & Hulin, C. L. (1973). An experimental evaluation of a temporal decay in the prediction of performance. *Organizational Behavior and Human Performance*, 9, 169-185.
- Anderson, J. E. (1939). The limitations of infant and preschool tests in the measure of intelligence. *Journal of Psychology*, 8, 351-379.
- Anderson, J. E. (1940). The prediction of terminal intelligence from infant and preschool tests. *Thirty-Ninth Yearbook, National Society for the Study of Education* (Pt. 1), 39, 385-403.
- Bechtoldt, H. P. (1960, April). *Statistical tests of predictions generated from factor hypotheses*. Paper presented at the Midwest Psychological Association, St. Louis.
- Bechtoldt, H. P. (1961). An empirical study of the factor analysis stability hypothesis. *Psychometrika*, 26, 405-432.
- Brenner, M. H., & Lockwood, H. C. (1965). Salary as a predictor of salary. *Journal of Applied Psychology*, 49, 295-298.
- Brogden, H. E., & Taylor, E. K. (1950). The dollar criterion—Applying the cost accounting concept to criterion construction. *Personnel Psychology*, 3, 133-154.
- Carlson, A. B., & Werts, C. E. (1976). *Relationships among law school predictors, law school performance, and bar examination results* (Report No. LSAC-76-1, Law School Admission Council Reports, Vol. 1, 1949-1969). Princeton, NJ: Law School Admission Council.
- Corballis, M. C. (1965). Practice and the simplex. *Psychological Review*, 72, 399-406.
- Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Chicago: University of Illinois Press.
- Dennis, W. (1954). Predicting scientific productivity in later maturity from records of earlier decades. *Journal of Gerontology*, 9, 465-467.
- Dennis, W. (1956). Age and productivity among scientists. *Science*, 123, 724-725.
- Dunham, R. B. (1974). Ability-skill relationships: An empirical explanation of change over time. *Organizational Behavior and Human Performance*, 12, 372-382.

- Fleishman, E. A. (1960). Abilities at different stages of practice in rotary pursuit performance. *Journal of Experimental Psychology*, 60, 162-171.
- Fleishman, E. A., & Hempel, W. E., Jr. (1954). Changes in factor structure of a complex psychomotor test as a function of practice. *Psychometrika*, 19, 239-252.
- Fleishman, E. A., & Hempel, W. E., Jr. (1955). The relationship between abilities and improvement with practice in a visual discrimination reaction task. *Journal of Experimental Psychology*, 49, 801-812.
- Fleishman, E. A., & Parker, J. F. (1959). *Prediction of advanced levels of proficiency in a complex tracking task* (Tech. Rep. No. WADC 59-255). Wright-Patterson Air Force Base, Ohio: Wright Air Development Center.
- Fleishman, E. H., & Rich, S. (1963). Role of kinesthetic and spatial-visual abilities in perceptual-motor learning. *Journal of Experimental Psychology*, 66, 6-11.
- Guttman, L. (1955). A generalized simplex for factor analysis. *Psychometrika*, 20, 173-192.
- Hobson v. Hansen, 269 F. Supp. 401 (D.D.C. 1967). Affirmed sub nom.
- Hulin, C. L., & Humphreys, L. G. (1980). Foundations of test theory: Construct validity in psychological measurement. *Proceedings of the Symposium on Theory and Application in Education and Employment* (pp. 5-12). Princeton, NJ: U.S. Office of Personnel Management and Educational Testing Service.
- Humphreys, L. G. (1960). Investigations of simplex. *Psychometrika*, 25, 313-323.
- Humphreys, L. G. (1968). The fleeting nature of the prediction of college academic success. *Journal of Educational Psychology*, 59, 375-380.
- Humphreys, L. G. (1985). *Intelligence: Three kinds of instability and their consequences for policy*. Unpublished manuscript, University of Illinois.
- Humphreys, L. G., & Davey, T. C. (1984). *Continuity in the development of intelligence from one year to 17*. Unpublished manuscript, University of Illinois.
- Humphreys, L. G., Davey, T. C., & Park, R. K. (1984). *A correlational analysis of standing height and intelligence*. Unpublished manuscript, University of Illinois.
- Humphreys, L. G., & Parsons, C. K. (1979). A simplex process model for describing differences between cross-legged correlations. *Psychological Bulletin*, 86, 325-334.
- Humphreys, L. G., & Taber, T. (1973). Postdiction study of the Graduate Record Examination and eight semesters of college grades. *Journal of Educational Measurement*, 10, 179-184.
- Hunter, J. E., & Hunter, R. F. (1984). Validity of utility of alternative predictors of job performance. *Psychological Bulletin*, 96, 72-98.
- Hunter, J. E., & Schmidt, F. L. (1982). Fitting people to jobs: The impact of personnel selection on national productivity. In M. D. Dunnette & E. A. Fleishman (Eds.), *Human performance and productivity: Human capability assessment* (pp. 233-284). Hillsdale, NJ: Erlbaum.
- Larry P. v. Riles, 495 F. Supp. 926 (N.D. Cal. 1979). Appeal docketed, No. 800-4027 (9th Cir. January 17, 1980).
- Lin, P. C., & Humphreys, L. G. (1977). Predictions of academic performance in graduate and professional school. *Applied Psychological Measurement*, 1, 249-257.
- PASE v. Hannon, 506 F. Supp. 831 (N.D. Ill. 1980).
- Perl, R. E. (1934). An application of Thurstone's method of factor analysis to practice series. *Journal of General Psychology*, 11, 209-212.
- Powers, D. E. (1982). Long-term predictive and construct validity of two traditional predictors of law school performance. *Journal of Educational Psychology*, 74, 468-576.
- Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). Impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology*, 66, 609-626.
- Siegel, B. (Ed.). (1985). *The Sporting News official baseball register*. St. Louis, MO: Sporting News Publishing Company.
- Smuck v. Hobson, 408 F. 2d 175 (D.C. Cir., 1969).
- Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection. *Journal of Applied Psychology*, 23, 565-578.
- Wilson, R. S. (1983). The Louisville twin study: Developmental synchronies in behavior. *Child Development*, 54, 298-316.

Received July 23, 1986

Revision received December 15, 1986

Accepted November 17, 1986 ■

Situational Specificity in Assessment Center Ratings: A Confirmatory Factor Analysis

Peter Bycio
Bowling Green State University

Kenneth M. Alvares
Frito-Lay Inc., Plano, Texas

June Hahn
Bowling Green State University

Assessment center ratings of eight abilities from each of five situational exercises were examined for their cross-situational consistency and discriminant validity. A series of confirmatory factor analyses revealed that the ratings were largely (if not totally) situation specific, and that assessors failed to distinguish among the eight target abilities. These results combined with previous research suggest that the assessment center method measures mainly situation-specific performance, not cross-situational managerial abilities. We suggest that the intended constructs might be better measured if more ability-related behaviors were elicited within each exercise and if the cognitive demands placed on assessors were reduced.

Most well-known applications of the assessment center method were designed in part to measure constructs assumed to represent conceptually distinct job-related abilities. For example, IBM's centers required assessors to use constructs such as oral communications, decision making, and administrative ability (Hinrichs, 1978). Typically, the task of the assessors has been to integrate observations of assessee behavior from a series of situational exercises, such as the *in-basket* and *group discussion*. These observations have then been used as a basis for ratings supposed to reflect the constructs of interest (Sackett & Dreher, 1982, 1984).

Noting that many centers were designed to measure complex abilities, several investigators (see, e.g., Sackett & Dreher, 1982; Turnage & Muchinsky, 1982) used the multitrait-multimethod matrix approach (cf. Campbell & Fiske, 1959) to evaluate the construct validity of assessment center ratings. This involved intercorrelating the ratings across exercises and examining the correlations to determine (a) the extent to which ratings of the same ability were similar across measurement methods (convergent validity) and (b) the extent to which ratings of different abilities were uncorrelated (discriminant validity). Convergent and discriminant validity help justify inferences regarding construct validity (Campbell & Fiske, 1959).

So far, the multitrait-multimethod matrix findings pertaining to assessment centers have not been positive. In centers studied by Sweeney (1976) and by Sackett and Dreher (1982), ratings of the same abilities were virtually uncorrelated across ex-

ercises. Although reasonably sized convergent validities have been found in some centers (Neidig & Neidig, 1984), pervasive halo, indicative of a lack of discriminant validity, has also typified the ratings from those programs (also see Silverman, Dalesio, Woods, & Johnson, 1986; Thomson, 1970; Turnage & Muchinsky, 1982). Multitrait-multimethod findings of this nature prompted Sackett and Dreher (1982) to conclude that there was "virtually no support for the view that the assessment center technique generated dimension scores that can be interpreted as representing complex constructs" (p. 409).

Unlike Sackett and Dreher (1982), Neidig and Neidig (1984) were not particularly surprised or alarmed by the poor or sometimes nonexistent evidence of convergent validity. They argued that properly designed situational exercises purposely place assesseees in a variety of job-related contexts, and therefore, "stable performance across exercises by all participants is not necessarily expected" (Neidig & Neidig, 1984, p. 184). Sackett and Dreher (1984) agreed that assessment center performance is "at least in part situationally determined" (p. 189), but questioned whether the method could then be used to measure broad cross-situational abilities.

An unresolved question seems to be the *extent* to which behavior is influenced by the situational context, as opposed to cross-situational abilities. This same question has been the focus of vigorous debate in the personality-social-psychology literature under the rubric of the cross-situational consistency versus cross-situational specificity issue (cf. Endler & Magnusson, 1976; Epstein, 1979, 1983; Epstein & O'Brien, 1985; Mischel & Peake, 1982, 1983).

In this investigation of an operational assessment center, we examined the cross-situational specificity and discriminant validity displayed by ratings of eight abilities from each of five situational exercises. Confirmatory factor analysis (CFA; Jöreskog & Sörbom, 1983), as recommended by Schmitt and Stults (1986), was used to evaluate the plausibility of three alternative

We gratefully acknowledge both John Arnold, who originally suggested using confirmatory factor analysis, and the very helpful comments of two anonymous reviews.

June Hahn is now with Procter and Gamble, Cincinnati, Ohio.

Correspondence concerning this article should be addressed to Peter Bycio, who is now at Xavier University, Department of Management, 3800 Victory Parkway, Cincinnati, Ohio 45207.

models of the assessor ratings. Each model implied the presence of different degrees of cross-situational consistency and discriminant validity.

Although CFA has been illustrated previously (see Kalleberg & Kluegel, 1975), a brief discussion of it is necessary to adequately communicate the hypotheses tested in this study. In particular, we carefully describe the rationale for the first model tested because the others are simply less general versions of the first.

Model 1: Eight Abilities and Five Exercises

The designers of the center intended that all eight abilities should be measured in each of the five exercises. The first model we evaluated reflected this premise. In terms of CFA, the implication was that the assessment center ratings should be describable by eight ability factors (one representing each ability), five exercise factors (one for each situational exercise), plus error variance. The form of this model is shown in Tables 1 and 2.

The factor matrix in Table 1 is different from those associated with traditional factor analysis in the sense that not all of the loadings are estimated. Certain loadings—in the present case, those thought to be small and insignificant—have been fixed at zero (0). Fixed loadings are not estimated when the CFA algorithm generates a solution; they must remain at their fixed value. Other loadings, which were hypothesized to be nonzero in the population were left free (?). Only these free loadings were estimated by the CFA algorithm. In sum, unlike traditional factor analysis, CFA works within the constraints reflected by free and fixed loadings to yield a factor structure that most closely models the observed correlation matrix of assessment center ratings.

The positioning of the free loadings reflects a particular set of hypotheses. For example, scanning Table 1 horizontally reveals that each rating has exactly three free loadings, one for the ability being measured, another for the exercise (situation) used, and a third for error. It is the relative size of each type of loading that bears directly on the situational specificity issue. Situational specificity arguments will be supported if the estimates for the free loadings on the ability factors turn out to be small and insignificant, whereas the ones on the exercise factors are large. Evidence for cross-situational consistency will be obtained for a given ability only if the loadings on the ability factor are large and significant across the exercises. By testing this particular model and comparing the size of the loadings, it is possible to identify the relative contribution that ability, situation, and error made to each assessment center rating (cf. Schmitt & Stults, 1986; Widaman, 1985; among others).

Table 2 shows the pattern of relations that were expected among the factors in the model. As was the case with the factor matrix (see Table 1), only those values that were expected to be significantly different from zero were left free to be estimated by the CFA algorithm (?). The correlations among the eight ability factors were left free because with past centers (see Thomson, 1970, among others) assessor judgments of different abilities have been highly correlated. The magnitude of these ability factor intercorrelations will be of particular interest as they relate

directly to discriminant validity. High correlations among ability ratings cast doubt on their construct validity (Campbell & Fiske, 1959).

The correlations among the exercise factors were also expected to be nonzero, and were therefore left free to be estimated (?). (See Table 2.) Nonzero correlations were hypothesized because the exercises were similar in certain respects; they each were examples of a common measurement method (the situational exercise) and were all intended to reflect the same job. The correlations among the ability and exercise factors were fixed at zero (0) so that each rating could be decomposed and interpreted in terms of independent sources of ability, exercise, and error variance (cf. Schmitt & Stults, 1986; Widaman, 1985). Data available from the first author revealed that the ability-exercise factor intercorrelations were usually less than .20 in this data set, and that allowing for these estimates produced negligible improvements in overall fit.

Model 2: One General Ability and Five Exercises

Factor Model 1 is only one of many plausible representations of the data; at least two others can be hypothesized from previous research. The rationale for Model 2 comes from the lack of discriminant validity found in some sets of assessment center ratings (see, e.g., Turnage & Muchinsky 1982). Given findings of this nature, it was reasonable to expect that poor discriminant validity would also characterize the present data set, making the eight separate ability factors shown in Table 1 unnecessary. Therefore, the second model we tested involved a single ability, five exercises, plus error. In other words, Model 2 was formed by collapsing the eight ability factors in the first model into one, reflecting the plausible expectation that the assessor ratings would not reflect eight separate abilities.

Model 3: Five Exercises

The third model involved removing all of the ability factors from Model 1, leaving only five exercise factors, plus error. This exercise-only model most closely resembled the findings of Sackett and Dreher (1982), who found that assessor ratings of the same ability were virtually uncorrelated across exercises in two of the three centers they studied.

Method

Ratings from a 1-day assessment center designed to meet selection and developmental objectives for the position of manufacturing supervisor were obtained from a large manufacturing firm. In all, 12 candidates per center—each recommended by his or her supervisor—were evaluated by assessors who held positions at least one level above them. The program was designed so that one assessor would observe candidate behavior in a given exercise. Each of five assessors saw a given candidate in a different exercise. The assessors met as a team after all of the exercises were completed to decide on overall ratings.

Subjects

A complete set of ratings from 1,170 candidates evaluated during the first 3 years of the program formed the basis of the analyses reported here.

Table 1
Factor Matrix Hypothesized to Underlie the Observed Assessment Center Ratings

Source of the rating	Ability factor								Exercise factor					Error factor (E)
	A1	A2	A3	A4	A5	A6	A7	A8	E1	E2	E3	E4	E5	
Problem-solving group														
Organizing-planning	?	0	0	0	0	0	0	0	?	0	0	0	0	?
Analyzing	0	?	0	0	0	0	0	0	?	0	0	0	0	?
Decision making	0	0	?	0	0	0	0	0	?	0	0	0	0	?
Controlling	0	0	0	?	0	0	0	0	?	0	0	0	0	?
Oral communications	0	0	0	0	?	0	0	0	?	0	0	0	0	?
Interpersonal relations	0	0	0	0	0	?	0	0	?	0	0	0	0	?
Influencing	0	0	0	0	0	0	?	0	?	0	0	0	0	?
Flexibility	0	0	0	0	0	0	0	?	?	0	0	0	0	?
In-basket														
Organizing-planning	?	0	0	0	0	0	0	0	0	?	0	0	0	?
Analyzing	0	?	0	0	0	0	0	0	0	?	0	0	0	?
Decision making	0	0	?	0	0	0	0	0	0	?	0	0	0	?
Controlling	0	0	0	?	0	0	0	0	0	?	0	0	0	?
Oral communications	0	0	0	0	?	0	0	0	0	?	0	0	0	?
Interpersonal relations	0	0	0	0	0	?	0	0	0	?	0	0	0	?
Influencing	0	0	0	0	0	0	?	0	0	?	0	0	0	?
Flexibility	0	0	0	0	0	0	0	?	0	?	0	0	0	?
Role-play														
Organizing-planning	?	0	0	0	0	0	0	0	0	0	?	0	0	?
Analyzing	0	?	0	0	0	0	0	0	0	0	?	0	0	?
Decision making	0	0	?	0	0	0	0	0	0	0	?	0	0	?
Controlling	0	0	0	?	0	0	0	0	0	0	?	0	0	?
Oral communications	0	0	0	0	?	0	0	0	0	0	?	0	0	?
Interpersonal relations	0	0	0	0	0	?	0	0	0	0	?	0	0	?
Influencing	0	0	0	0	0	0	?	0	0	0	?	0	0	?
Flexibility	0	0	0	0	0	0	0	?	0	0	?	0	0	?
Human relations group														
Organizing-planning	?	0	0	0	0	0	0	0	0	0	0	?	0	?
Analyzing	0	?	0	0	0	0	0	0	0	0	0	?	0	?
Decision making	0	0	?	0	0	0	0	0	0	0	0	?	0	?
Controlling	0	0	0	?	0	0	0	0	0	0	0	?	0	?
Oral communications	0	0	0	0	?	0	0	0	0	0	0	?	0	?
Interpersonal relations	0	0	0	0	0	?	0	0	0	0	0	?	0	?
Influencing	0	0	0	0	0	0	?	0	0	0	0	?	0	?
Flexibility	0	0	0	0	0	0	0	?	0	0	0	?	0	?
Interview														
Organizing-planning	?	0	0	0	0	0	0	0	0	0	0	0	?	?
Analyzing	0	?	0	0	0	0	0	0	0	0	0	0	?	?
Decision making	0	0	?	0	0	0	0	0	0	0	0	0	?	?
Controlling	0	0	0	?	0	0	0	0	0	0	0	0	?	?
Oral communications	0	0	0	0	?	0	0	0	0	0	0	0	?	?
Interpersonal relations	0	0	0	0	0	?	0	0	0	0	0	0	?	?
Influencing	0	0	0	0	0	0	?	0	0	0	0	0	?	?
Flexibility	0	0	0	0	0	0	0	?	0	0	0	0	?	?

Note. As implied by the hypothesis, the 0s are fixed parameters, whereas the ?s represent those loadings left free to be estimated.

Situational Exercises

Problem-solving group exercise. Each of three candidates was presented with a problem situation and was asked to rank the desirability of alternative solutions to it. After a group discussion, the alternatives were reranked, this time by the assesseees as a group.

24-item in-basket. The 24 items in the in-basket exercise were similar to those faced by a manufacturing supervisor. Assesseees were given 1 hr to complete the in-basket. The assessor then conducted a 45-min in-basket interview to explore the candidate's approach to each item.

Role-play exercise. For 1 hr, assesseees dealt with four different problem situations typically faced by a manufacturing supervisor. For exam-

ple, one situation demanded that the candidate explain the assembly of a product to the assessor.

Human relations group exercise. For 50 min an assessor observed candidates discussing the firing of a fictitious employee.

Interview. A 30-min interview was structured around assessee responses to a personal information questionnaire.

Abilities Measured

In each exercise, 9-point scales, ranging from *extremely low* (1) to *extremely high* (9), were used to obtain ratings on each of the following

Table 2
Hypothesized Pattern of Factor Intercorrelations

Factor	A1	A2	A3	A4	A5	A6	A7	A8	E1	E2	E3	E4	E5
A1. Organizing and planning	—												
A2. Analyzing	?	—											
A3. Decision making	?	?	—										
A4. Controlling	?	?	?	—									
A5. Oral communications	?	?	?	?	—								
A6. Interpersonal relations	?	?	?	?	?	—							
A7. Influencing	?	?	?	?	?	?	—						
A8. Flexibility	?	?	?	?	?	?	?	—					
E1. Problem solving	0	0	0	0	0	0	0	0	—				
E2. In-basket	0	0	0	0	0	0	0	0	?	—			
E3. Role play	0	0	0	0	0	0	0	0	?	?	—		
E4. Human relations	0	0	0	0	0	0	0	0	?	?	?	—	
E5. Interview	0	0	0	0	0	0	0	0	?	?	?	?	—

Note. As implied by the hypothesis, the 0s are fixed parameters, whereas the ?s represent those correlations left free to be estimated.

eight abilities: organizing and planning, analyzing, decision making, controlling, oral communications, interpersonal relations, influencing, and flexibility.

Statistical Analyses

The assessment center ratings were intercorrelated and displayed in a 40 × 40 (8 abilities, each measured by 5 exercises) correlation matrix. Using the LISREL VI software (Jöreskog & Sörbom, 1983), CFA was used to test the plausibility of the three models described earlier.

LISREL VI generates an estimated correlation matrix using the hypothesized factor structure as a guide. If only small differences exist between the actual and the estimated matrices, then the hypothesized factor structure is thought to be a plausible one. Following Harvey, Billings, and Nilan (1985), four indices were used to assess the similarity between the two matrices and to help evaluate which of the three hypothesized models was the best representation of the assessment center ratings. The four indices are (a) root mean square residual (Jöreskog & Sörbom, 1983), wherein small values imply a good fit because it reflects the size of the differences obtained when the estimated correlation matrix is subtracted from the observed one; (b) rho (Bentler & Bonett, 1980), wherein values of .90 or higher reflect a reasonable model because it compares the fit of the particular model of interest against the null model—in this case, one that hypothesizes that all the assessment center ratings are uncorrelated in the population; (c) parsimonious fit index (James, Mulaik, & Brett, 1982), which, like rho, uses the null model as a standard against which to evaluate alternative models, but also incorporates the differences in the degrees of freedom associated with the models involved; and (d) chi-square (Jöreskog & Sörbom, 1983), which although not directly interpreted, was incorporated into both rho and the parsimonious fit index.

Results

Descriptive statistics for the assessment center ratings are shown in Table 3. The means of the ability ratings ranged from 5.18 (in-basket rating of influencing) to 5.98 (interview rating of oral communications), whereas the standard deviations were between 1.17 (problem-solving group rating of interpersonal relations) and 1.74 (in-basket rating of organizing and planning). The Kolomogorov *D* statistic (Statistical Analysis System, 1982, p. 580) was used to test the hypothesis that these ratings

were a sample from a population with a normal distribution. This hypothesis was rejected in every case; the ratings were negatively skewed. This means that the chi-squares, *T* values, and standard errors from LISREL must all be viewed with caution because they were calculated under the assumption of multivariate normality.

Table 4 shows the observed correlation matrix. The average correlation involving ratings of the same ability from two different exercises was .36, whereas the average correlation of ability ratings within an exercise was .75.

The fit indices associated with each model are shown in Table 5. One of the models not specifically addressed earlier (5 abilities and 5 exercises) was evaluated because we found that the eight abilities–five exercises model produced correlations among certain abilities that were slightly greater than 1. The appearance of technically impossible correlations (such as 1.02) reminds one that LISREL produces estimates of parameters, not exact values. It also indicated that eight separate abilities were not reflected in the data. To remove the perfect correlations, we formed the five abilities–five exercises model (hereafter referred to as Model 1') by combining (a) organizing and planning, and analyzing, (b) controlling and influencing, and (c) interpersonal relations and flexibility.

To decide which of the models was the best representation of the data, a statistical comparison involving the differences among the respective chi-squares is possible (Long, 1983). Unfortunately, the test is not very informative with large sample sizes because it then has the power to detect even very small differences between a model and the data. When the sample size is large, the chi-square will almost always reject a given model as inadequate, even if it is a good representation of the data in practical terms (Harvey et al., 1985; James et al., 1982).

Barring the direct use of the chi-square, several nonstatistical criteria for comparing the fit of nested models have been developed. For example, Bentler and Bonett (1980) argued that a model reasonably represents the observed data when rho is at least .90. As Table 5 shows, all of the models meet this criterion. Each model (with the possible exception of the 5 exercises representation) also has a small root mean square residual.

Table 3
Descriptive Statistics for the Assessment Center Ratings

Source of the rating	<i>M</i>	<i>SD</i>	Skewness	Kurtosis	<i>D</i> ^a	<i>p</i> < <i>D</i> ^a
Problem-solving group						
Organizing and planning	5.42	1.20	-.31	0.09	.17	.01
Analyzing	5.65	1.25	-.54	0.67	.18	.01
Decision making	5.56	1.27	-.44	0.38	.17	.01
Controlling	5.55	1.41	-.40	0.28	.15	.01
Oral communications	5.86	1.20	-.42	0.43	.18	.01
Interpersonal relations	5.77	1.17	-.32	0.40	.19	.01
Influencing	5.56	1.41	-.46	0.37	.16	.01
Flexibility	5.58	1.22	-.53	0.67	.19	.01
In-basket						
Organizing and planning	5.25	1.74	-.10	0.62	.11	.01
Analyzing	5.23	1.69	-.15	-0.55	.13	.01
Decision making	5.34	1.55	-.21	-0.26	.14	.01
Controlling	5.20	1.51	-.24	-0.10	.14	.01
Oral communications	5.70	1.44	-.28	0.12	.14	.01
Interpersonal relations	5.60	1.32	-.26	0.37	.15	.01
Influencing	5.18	1.42	-.28	0.24	.16	.01
Flexibility	5.45	1.42	-.33	0.09	.15	.01
Role play						
Organizing and planning	5.79	1.40	-.35	-0.07	.16	.01
Analyzing	5.76	1.42	-.42	0.00	.16	.01
Decision making	5.83	1.35	-.36	-0.01	.16	.01
Controlling	5.71	1.42	-.38	0.02	.16	.01
Oral communications	5.96	1.35	-.29	-0.01	.14	.01
Interpersonal relations	5.88	1.22	-.27	0.40	.16	.01
Influencing	5.65	1.39	-.38	-0.12	.16	.01
Flexibility	5.79	1.35	-.47	0.21	.17	.01
Human relations group						
Organizing and planning	5.53	1.18	-.56	1.11	.18	.01
Analyzing	5.68	1.27	-.47	0.66	.18	.01
Decision making	5.69	1.26	-.70	1.03	.21	.01
Controlling	5.48	1.40	-.41	0.44	.15	.01
Oral communications	5.78	1.27	-.48	0.80	.18	.01
Interpersonal relations	5.69	1.21	-.45	0.95	.19	.01
Influencing	5.47	1.41	-.49	0.39	.16	.01
Flexibility	5.52	1.24	-.47	0.74	.17	.01
Interview						
Organizing and planning	5.73	1.27	-.27	0.30	.17	.01
Analyzing	5.71	1.27	-.41	0.48	.18	.01
Decision making	5.78	1.31	-.43	0.30	.18	.01
Controlling	5.65	1.35	-.34	0.21	.16	.01
Oral communications	5.98	1.40	-.46	0.39	.16	.01
Interpersonal relations	5.95	1.21	-.49	0.79	.18	.01
Influencing	5.64	1.34	-.37	0.28	.17	.01
Flexibility	5.76	1.24	-.33	0.21	.18	.01

Note. *N* = 1,170. ^a Kolomogorov *D* statistic.

Widaman (1985) argued that differences in rho of .01 or more were important on practical grounds. Using this standard, the five ability-five exercises model would be preferred over the exercise-only model, but not over the one ability-five exercises model. If the comparison were based on the one ability-five exercises model, then the exercises-only model would be preferred (cf. Widaman, 1985).

Another criterion that has been used to decide among models is the parsimonious fit index (see James et al., 1982). Parsimony is of central importance here. One seeks the model with the most degrees of freedom (i.e., the fewest freely estimated parameters) that still provides a reasonable representation of the

data. Comparison of the parsimonious fit indexes in Table 5 reveals that the exercises-only model meets this criterion.

We could not unambiguously choose among the alternative models. Although the exercises-only model was parsimonious, it implied a complete lack of cross-situational consistency in behavior that belied the average convergent validity of .36 in the data. The exercises-only model also had 33 large (greater than 2) normalized residuals, indicative of many specification errors (cf. Jöreskog & Sörbom, 1983). Furthermore, the pattern of the large residuals was not random. They were primarily confined to a few variables (in particular, oral communications).

Table 4
Observed Correlations Among the Assessment Center Ratings

Source of the rating	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Problem-solving group															
1. Organizing and planning	—														
2. Analyzing	.77	—													
3. Decision making	.76	.78	—												
4. Controlling	.74	.76	.79	—											
5. Oral communications	.67	.72	.71	.72	—										
6. Interpersonal relations	.64	.67	.68	.69	.69	—									
7. Influencing	.71	.77	.80	.83	.71	.70	—								
8. Flexibility	.67	.71	.71	.70	.67	.73	.73	—							
In-basket															
9. Organizing and planning	.32	.33	.33	.32	.32	.26	.30	.27	—						
10. Analyzing	.35	.36	.34	.35	.33	.28	.33	.29	.87	—					
11. Decision making	.31	.33	.34	.33	.32	.26	.32	.27	.82	.86	—				
12. Controlling	.35	.35	.36	.36	.33	.28	.35	.30	.82	.83	.84	—			
13. Oral communications	.33	.33	.34	.35	.37	.28	.34	.31	.70	.71	.73	.75	—		
14. Interpersonal relations	.30	.28	.30	.30	.31	.27	.29	.27	.65	.66	.69	.70	.70	—	
15. Influencing	.34	.34	.33	.35	.34	.29	.33	.29	.79	.81	.82	.85	.77	.72	—
16. Flexibility	.33	.33	.33	.34	.33	.27	.32	.27	.76	.79	.80	.79	.74	.73	.82
Role play															
17. Organizing and planning	.33	.34	.35	.33	.34	.28	.32	.32	.40	.42	.41	.38	.36	.33	.41
18. Analyzing	.35	.36	.36	.33	.36	.29	.33	.34	.38	.41	.40	.37	.36	.32	.41
19. Decision making	.34	.34	.36	.34	.35	.30	.33	.32	.34	.36	.36	.34	.34	.30	.38
20. Controlling	.36	.37	.39	.36	.38	.33	.36	.35	.34	.36	.36	.36	.35	.30	.39
21. Oral communications	.37	.38	.40	.37	.44	.35	.38	.36	.38	.38	.39	.37	.43	.33	.40
22. Interpersonal relations	.28	.27	.30	.28	.31	.28	.29	.28	.30	.31	.30	.31	.30	.28	.34
23. Influencing	.37	.37	.39	.39	.39	.33	.37	.35	.38	.40	.40	.40	.40	.33	.43
24. Flexibility	.31	.32	.33	.32	.33	.28	.30	.30	.35	.37	.38	.36	.36	.33	.40
Human relations group															
25. Organizing and planning	.41	.39	.39	.38	.39	.31	.37	.35	.34	.34	.34	.36	.36	.32	.38
26. Analyzing	.39	.38	.38	.38	.42	.31	.37	.36	.34	.37	.35	.38	.39	.32	.40
27. Decision making	.39	.38	.38	.38	.41	.31	.37	.35	.33	.35	.33	.36	.36	.30	.39
28. Controlling	.37	.35	.37	.38	.40	.29	.36	.32	.33	.36	.36	.37	.38	.31	.40
29. Oral communications	.29	.40	.40	.40	.45	.33	.39	.36	.34	.36	.34	.37	.42	.31	.40
30. Interpersonal relations	.36	.34	.35	.34	.40	.32	.33	.34	.29	.31	.28	.31	.34	.30	.34
31. Influencing	.39	.39	.39	.39	.41	.30	.37	.35	.33	.36	.34	.37	.38	.30	.40
32. Flexibility	.36	.36	.36	.35	.40	.30	.35	.33	.33	.35	.34	.36	.38	.32	.38
Interview															
33. Organizing and planning	.32	.32	.35	.32	.37	.27	.33	.29	.38	.38	.39	.42	.42	.37	.42
34. Analyzing	.33	.33	.37	.33	.37	.27	.33	.32	.35	.35	.37	.41	.41	.33	.40
35. Decision making	.34	.33	.36	.34	.38	.27	.34	.32	.33	.34	.37	.39	.39	.33	.38
36. Controlling	.32	.32	.35	.35	.38	.28	.33	.31	.34	.34	.36	.37	.38	.32	.37
37. Oral communications	.38	.37	.42	.41	.46	.33	.40	.35	.33	.34	.36	.39	.46	.33	.39
38. Interpersonal relations	.27	.26	.30	.27	.33	.26	.27	.28	.29	.26	.28	.32	.35	.29	.31
39. Influencing	.32	.32	.35	.34	.37	.27	.34	.30	.34	.34	.36	.38	.40	.34	.38
40. Flexibility	.30	.30	.32	.31	.35	.27	.31	.30	.30	.30	.32	.35	.37	.32	.35

Note. Decimal points and the diagonal elements have been omitted.

The one ability–five exercises model also had conceptual problems. Five of the ability loadings were estimated to be zero. This was inconsistent with the observed correlations, wherein every convergent validity was significantly different from zero. As with the exercises-only representation, there were also a fair number (13) of large normalized residuals associated, again, primarily with a few of the variables.

Finally, although the five ability–five exercises model had the largest rho (.976), a small root mean square residual (.026), and the fewest large normalized residuals (9), it also had the greatest number of free loadings to estimate. A better fit would be expected on that basis alone (cf. James et al., 1982).

Rather than make a somewhat arbitrary choice among the

competing models, we present condensed results for all three. (Findings for Model 1' are shown as opposed to those for Model 1.) Table 6 shows all of the squared, freely estimated factor loadings for each model. These can be interpreted as variance components reflecting the extent to which each assessment center rating was influenced by ability, exercise, and error (Schmitt & Stults, 1986; Widaman, 1985).

Table 6 displays an important trend. Irrespective of the particular model involved, exercise variance dominates these ratings. Across the three models, the exercise contributed more variance than did ability and error combined. For example, with Model 1' (5 abilities–5 exercises), the average ability variance was only .16. The equivalent figure for Model 2 (1 ability–

16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
—																								
41	—																							
40	83	—																						
36	80	81	—																					
37	79	78	82	—																				
39	70	70	73	74	—																			
32	64	67	68	68	72	—																		
41	76	78	79	81	75	73	—																	
38	74	75	76	76	70	71	81	—																
36	36	36	36	35	39	30	39	35	—															
36	39	39	39	40	43	33	43	39	75	—														
36	37	39	39	39	42	31	43	38	72	79	—													
36	34	35	36	35	39	29	39	33	72	76	77	—												
37	35	36	37	37	43	33	41	37	70	75	74	76	—											
31	30	30	31	31	36	30	34	30	69	73	71	71	75	—										
36	36	36	36	36	41	30	40	35	71	78	79	84	75	70	—									
37	36	36	36	36	40	31	39	36	69	75	73	72	72	77	76	—								
40	40	41	41	39	44	34	42	40	42	42	41	42	44	38	42	40	—							
36	37	39	38	37	40	30	40	37	40	40	38	39	42	34	40	36	84	—						
36	37	37	38	36	38	28	39	37	36	37	37	38	41	32	39	36	81	83	—					
36	35	37	37	36	39	31	38	36	37	36	36	38	40	35	37	37	79	79	83	—				
37	37	40	40	39	45	33	39	39	39	42	40	41	49	38	42	39	75	75	76	76	—			
31	30	32	33	32	32	27	33	32	29	29	31	27	32	28	30	30	69	71	72	72	72	—		
37	36	37	39	37	40	31	39	38	37	37	38	38	40	32	39	35	79	80	81	83	78	75	—	
34	33	35	36	36	37	29	37	37	34	34	35	35	36	31	35	33	76	77	78	78	75	75	80	—

5 exercises) was .04. Model 3 (5 exercises) implies that there is zero systematic ability variance.

Table 7 shows the estimated correlations among the factors for the five ability–five exercises model. The correlations among the five ability factors were very high, ranging from .84 (between oral communications and the organizing–planning–analyzing variables) to .98 (between the decision-making and the controlling–influencing variables). Although all of the ability factor correlations were more than two standard errors from unity (making them less than perfectly correlated in the population), they were, nonetheless, very high. The results for the other models (see Table 8) imply the same conclusion because neither of them contain factors representing more than one ability.

Moderate correlations were observed among the exercise fac-

tors in all three models (see Tables 7 and 8). The average exercise factor intercorrelation was .36 for the five abilities–five exercises model, .44 for the one ability–five exercises model, and .47 for the exercise-only representation.

Discussion

Estimates from each of three models indicated that the assessment center ratings were largely situation specific. Also, when hypotheses concerning discriminant validity were tested, perfect or near-perfect correlations were observed among many of the ability factors. Although there may still be evidence in these findings for a form of behavioral consistency referred to as coherence (cf. Magnusson & Endler, 1977), the ratings did

Table 5
Fit Indices for Models of the Assessment Center Ratings

Model	χ^2	df	rho	Parsimonious fit index	Root mean square residual
Null	51602.85	780	—	—	.441
8 abilities–5 exercises	1662.98	662	.977	.821	.027
5 abilities–5 exercises	1745.09	680	.976	.842	.026
1 ability–5 exercises	2075.59	690	.969	.849	.026
0 abilities–5 exercises	2590.35	730	.961	.889	.031

not appear to reflect the eight broad cross-situational abilities that the program was designed to measure.

The results from this center are similar to those of previous studies. With the possible exception of Silverman et al. (1986), the evidence for cross-situational consistency (convergent validity) has been poor (cf. Turnage & Muchinsky, 1982) to nonexistent (cf. Sackett & Dreher, 1982). The same can be said of the data pertaining to discriminant validity (cf. Silverman et al., 1986; Thomson, 1970; Turnage & Muchinsky, 1982). Although broad generalizations about the assessment center method cannot be made with certainty (Turnage & Muchinsky, 1984), the preponderance of evidence suggests that the method does not measure large sets of job-related abilities.

Hypotheses Concerning the Exercise Effects

It has been suggested previously that assessment center behavior is situation specific (Neidig & Neidig, 1984; Sackett & Dreher, 1984). The models presented here documented the extent to which exercise variance pervades. This variance cannot simply be attributed to rater differences because, although single assessors made the exercise ratings, the program was designed so that exercises and assessors were uncorrelated. In other words, with the exception of the coordinator (who was exclusively responsible for ratings from the human relations group exercise), assessors rotated equally among the exercises during each run of the program. Furthermore, interrater reliability estimates available from a previous, similar program were in the high .60s to low .70s. These reliabilities correspond well with the error component estimates from the present program (see Table 6), which were usually about .30. Thus, rater differences were an unlikely source of situation variance in this study.

Speculation as to the nature of the exercise variance is prompted by the correlations among the exercise factors. They were not high considering that the same measurement method (the situational exercise) and the same job were involved. Their magnitude suggests that the exercises differed on some important parameter(s) that perhaps limited the extent to which cross-situational consistency could be demonstrated. For example, there were structural differences among exercises, including whether candidates were observed and judged as individuals (interview exercise) or as part of a group (human relations group exercise). The situations also differed structurally in that some ability judgments were made using written material (in-basket), whereas others were based on verbal behavior (role-play

exercise). Other less salient structural differences included the number of assessees involved, the length of the exercise, and the extent to which the assessor directly interacted with candidates.

Even when exercises are similar in structure, they differ in content (Neidig & Neidig, 1984). Our findings suggest that content differences might be pivotal. For example, the correlation between the two structurally similar group discussion exercises was, at most, .50 (depending on the model). Although both elicited verbal behavior in a group setting, they shared, at most, only 25% exercise variance. One crucial remaining difference between the two situations was discussion content.

The low to moderate correlations among even structurally similar situational exercises raises some interesting implications for the design of assessment centers. For example, we agree with Neidig and Neidig (1984) that the accurate sampling of situations is crucial to establishing the job relatedness of the method, but wonder along what lines sampling should proceed, given the apparently pervasive situational influence on behavior. Specifically, how exactly does one know when the most important situational aspects of a job have been reflected by a set of exercises? Because an assessment center cannot possibly model all of the situational aspects of a target position, exercise designers are essentially left guessing. There has been lack of systematic study of situations and their impact on behavior (Endler & Magnusson, 1976).

An example will illustrate the scope of the problem. Consider the manager who is being assessed for a possible promotion. Further, imagine that this person happens to be a poor speller, with illegible handwriting. At work, these problems are overcome through the use of a word processor and a spelling check program. But the assessment center has no word processor. As a result, the in-basket from this manager is difficult to read. It is full of spelling errors. The candidate also stumbles through the oral presentation exercise because the presentation material has been written in long hand, not neatly typed. This manager, to the surprise of his superiors, does not do well at the center and is passed up for the promotion.

Should all centers be equipped with word processors? More generally, to what extent must the exercises exactly reflect the job situation? Furthermore, is there such a thing as *the* job situation? In other words, is behavior on the job as situation specific as it appears to be in the assessment center context? If it is, performance appraisal systems geared to the measurement of managerial abilities may need to be scrapped in favor of situationally based judgments.

Table 6
Variance Components Associated With Models (M) 1', 2, and 3

Source of the rating	Variance component								
	Ability			Exercise			Error		
	M1'	M2	M3	M1'	M2	M3	M1'	M2	M3
Problem-solving group									
Organizing and planning	.11	.02	—	.62	.68	.70	.30	.30	.30
Analyzing	.09	.01	—	.70	.76	.76	.23	.23	.24
Decision making	.10	.02	—	.71	.77	.79	.21	.21	.21
Controlling	.10	.03	—	.71	.75	.79	.20	.21	.21
Oral communications	.13	.08	—	.56	.60	.67	.31	.32	.33
Interpersonal relations	.06	.05	—	.59	.58	.62	.36	.37	.38
Influencing	.10	.04	—	.72	.76	.80	.20	.21	.20
Flexibility	.07	.03	—	.62	.63	.66	.33	.34	.34
In-basket									
Organizing and planning	.10	.00	—	.72	.81	.79	.17	.18	.21
Analyzing	.09	.00	—	.80	.86	.83	.10	.13	.17
Decision making	.16	.00	—	.67	.84	.84	.16	.16	.16
Controlling	.24	.01	—	.59	.83	.84	.16	.16	.16
Oral communications	.41	.12	—	.36	.63	.67	.22	.25	.33
Interpersonal relations	.33	.06	—	.33	.56	.59	.34	.38	.41
Influencing	.31	.03	—	.53	.80	.83	.15	.17	.17
Flexibility	.27	.02	—	.53	.75	.77	.20	.22	.23
Role play									
Organizing and planning	.09	.00	—	.69	.80	.77	.21	.20	.23
Analyzing	.10	.00	—	.70	.80	.78	.21	.20	.22
Decision making	.09	.01	—	.73	.80	.80	.19	.20	.20
Controlling	.09	.02	—	.73	.78	.80	.19	.20	.20
Oral communications	.16	.07	—	.56	.63	.68	.29	.30	.32
Interpersonal relations	.08	.05	—	.56	.57	.61	.38	.38	.39
Influencing	.13	.03	—	.69	.77	.81	.19	.19	.19
Flexibility	.11	.03	—	.65	.71	.74	.25	.26	.26
Human-relations group									
Organizing and planning	.17	.03	—	.52	.65	.68	.31	.32	.32
Analyzing	.16	.04	—	.63	.74	.78	.21	.22	.22
Decision making	.13	.03	—	.64	.73	.77	.23	.23	.23
Controlling	.15	.03	—	.65	.75	.78	.21	.22	.22
Oral communications	.17	.09	—	.58	.65	.73	.24	.25	.27
Interpersonal relations	.12	.07	—	.60	.62	.69	.29	.31	.31
Influencing	.14	.03	—	.67	.77	.80	.20	.20	.20
Flexibility	.13	.05	—	.62	.67	.72	.27	.28	.28
Interview									
Organizing and planning	.27	.04	—	.57	.75	.79	.18	.21	.21
Analyzing	.26	.04	—	.59	.77	.81	.17	.19	.19
Decision making	.19	.03	—	.66	.79	.82	.17	.17	.18
Controlling	.13	.04	—	.70	.77	.81	.18	.19	.19
Oral communications	.23	.13	—	.54	.62	.73	.22	.25	.27
Interpersonal relations	.12	.07	—	.57	.59	.65	.32	.34	.35
Influencing	.16	.06	—	.69	.76	.82	.17	.18	.18
Flexibility	.15	.07	—	.65	.69	.76	.22	.24	.24

Note. M1' is the five abilities–five exercises model; M2 is the one ability–five exercises Model; M3 is the exercises-only representation.

The Search For Cross-Situational Consistency

Is it time to abandon the concept of broad cross-situational managerial abilities? The consensus in the personality–social-psychology literature is that behavior is influenced by *both* the situation and cross-situational consistency (Epstein, 1983; Mischel & Peake, 1983), but that the consistencies are more difficult to measure (Kenrick & Stringfield, 1980; Epstein, 1979, 1983). Kenrick and Stringfield (1980), for example, found that evidence for cross-situational consistency was easier to obtain

when the target traits were highly publicly observable. Perhaps this explains why oral communications—which in our opinion was among the most easily observable—had larger ability variances (in both Models 1' and 2) than any of the other target abilities.

In the personality–social-psychology literature, Epstein (1979, 1983) has also developed an idea with possible relevance to the weak relation among ability ratings from different exercises. He argued that one cannot hope to obtain high correlations between single instances of behavior from two different

Table 7
LISREL Estimates of the Factor Intercorrelations for Model 1'

Factor	A1	A2	A3	A4	A5	E1	E2	E3	E4	E5
A1. Organizing–planning–analyzing	—									
A2. Decision making	.96	—								
A3. Controlling–influencing	.92	.98	—							
A4. Oral communications	.84	.87	.89	—						
A5. Interpersonal–relations–flexibility	.87	.91	.93	.89	—					
E1. Problem solving	.00	.00	.00	.00	.00	—				
E2. In-basket	.00	.00	.00	.00	.00	.32	—			
E3. Role play	.00	.00	.00	.00	.00	.39	.38	—		
E4. Human relations	.00	.00	.00	.00	.00	.42	.32	.40	—	
E5. Interview	.00	.00	.00	.00	.00	.35	.30	.39	.37	—

Note. Model 1' = five abilities–five exercises.

situations because situational uniqueness masks the consistency. According to Epstein (1979, 1983), response consistencies emerge only when situational effects have been canceled out by averaging behavior over a sufficient number of occasions. To illustrate, he had college students keep records of various aspects of their behavior on each of 12 days. He then calculated the mean of the ratings from the even days and correlated it with the mean of the observations from the odd days. The resulting cross-situational correlations between the odd and even days ranged from .70 to .94 (Epstein, 1983, p. 181).

Sackett and Dreher (1984) indirectly tested the relevance of the Epstein (1979) logic to assessment centers. They treated all of the exercise-specific ratings of an ability as single items and then calculated coefficient alphas among the ratings of the same ability across exercises. The alphas ranged from $-.06$ (for written communication) to $.65$ (for oral communication), indicating that the addition of more exercises may have enhanced the evidence for cross-situational consistency for some of the abilities. Alphas were also calculated for ratings within exercises. These were much higher ($M = .90$) than those involving cross-situational assessments.

Sackett and Dreher (1984) concluded that the alphas also reflected the predominance of situation variance, and suggested that assessment centers be geared to measuring performance in situations, as opposed to cross-situational abilities. We agree that this is likely to be the most practical solution, but would also like an empirical test focused on improving the informa-

tion level within existing exercises, as opposed to one geared simply to adding a greater number of typical simulations to an already existing program. In other words, unlike Epstein (1979, 1983), who aggregated observations over a wide range of situations, we wish to highlight the potential importance of obtaining multiple ability-related observations within each exercise. By providing assessors with a large number of behaviors to intuitively aggregate within each exercise, high correlations across exercises may result.

Does the typical situational exercise allow assessee the opportunity to repeatedly exhibit ability-related behavior? In some cases we think not. Note that before aggregation of any sort can occur, assessors must observe at least two behaviors for each target ability. If the program was designed (as this one was) to measure eight different abilities, each exercise must elicit a minimum of 16 ability-relevant behaviors before aggregation is even a possibility. In our experience, most situational exercises do not reliably elicit anywhere near this number of behaviors. Instead, assessors within an exercise are sometimes (if not usually) forced to base all of their judgments on four or five behaviors. Extreme examples come to mind of a candidate who contributes one or two sentences during an entire group discussion, or of the individual who spends all of the allotted time focusing on a single issue in a performance appraisal.

One typically used assessment center exercise that has the potential to elicit a large number of ability-related behaviors is the in-basket. In this respect, it is not surprising that the esti-

Table 8
LISREL Estimates of the Factor Intercorrelations for Models 2 and 3

Factor	Model 2: Five exercises–one ability						Model 3: Five exercises only				
	A1	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5
A1. General ability	—										
E1. Problem solving	.00	—					.42	—			
E2. In-basket	.00	.41	—				.46	.48	—		
E3. Role play	.00	.44	.47	—			.50	.46	.49	—	
E4. Human relations	.00	.47	.45	.47	—		.44	.46	.48	.49	—
E5. Interview	.00	.40	.45	.46	.45	—					

mates from Model 1' indicate that the in-basket (and subsequent oral interview) produced ratings with the highest proportion of ability variance for five of the eight originally targeted abilities. This finding was unique to Model 1', however, and it is important to note that even a 24-item in-basket does not ensure that a large number of behaviors relevant to each ability will be elicited. The in-basket usually contains junk items and interrelated documents that cut down on the number of potential observations. Tight time limits, although useful for allowing judgments about prioritizing, can also leave assessors in a bind when, for example, a candidate meticulously completes 4 items but leaves the rest blank for a lack of time.

Cognitive Demands on Assessors

Even if exercises were developed so that a large number of ability-relevant behaviors were reliably elicited, we cannot assume that the assessors could observe, record, and aggregate them all. Shack (1983), using findings from the cognitive literature, suggested that basic limitations in human information processing make it very difficult for assessors to observe and record behaviors during assessment centers (also see Silverman et al. 1986, for work following from Shack, 1983). Rather than deal with all of the behaviors from an exercise, Shack (1983) argued that the assessors probably attend to some and ignore others. If true, this further reduces the likelihood that each of the exercise-based ability ratings will reflect enough behavioral observations to be cross situational. Indeed, some of the exercises from the present center placed incredible demands on the observation and recording skills of assessors. In the problem-solving group simulation, for example, a single assessor was responsible for observing and recording the behavior of three candidates over a 45-min period. In the human relations group exercise, an individual assessor was asked to spend 50 min observing and recording the behavior of four candidates. These not-so-atypical cognitive demands would probably make it almost impossible for assessors to observe and record enough behavior on which to base eight ability ratings for each candidate, even assuming enough ability-related behaviors were elicited in the first place.

The task of the assessors would also be simpler if there were fewer target abilities. Discriminant validity might improve as a result (Silverman et al., 1986; Turnage & Muchinsky, 1982, among others), and such a reduction would make it more practical to ensure that each exercise yielded enough behavior on which to judge the complete set of abilities.

Conclusion

We have emphasized that assessment center ratings usually do not reflect all of the intended constructs. Yet, in their meta-analysis, published as a monograph in this issue, Gaugler, Rosenthal, Thornton, and Bentson (1987) concluded that criterion-related validities for the assessment center method were typically nonzero and generalizable. For some, these criterion-related findings make the less than impressive multitrait-multimethod evidence easy to overlook. For example, Neidig and Neidig (1984) argued that positive multitrait-multimethod evi-

dence is not necessary to demonstrate that assessment center ratings are job related. They cited evidence from Borman (1982), in which ratings that apparently reflected exercise variance, nonetheless exhibited criterion-related validity.

We would not be surprised if the ratings from this study also exhibit validity in the criterion-related sense. The exercise-specific ratings would be expected to correlate with job performance to the degree that the exercise involved accurately reflected the job itself. It is also possible that the assessors in this program used the consensus meeting to intuitively aggregate information from across the exercises and to generate overall ratings that will meet the standards for criterion-related validity. Indeed, when the program is designed exclusively for selection purposes, validity in this limited sense may be all that is necessary.

Many programs, though, are used—as this one was—not only for selection, but also for managerial development. When used as a developmental tool, the method is supposed to provide candidates with feedback concerning their managerial strengths and weaknesses. It is in this context that positive multitrait-multimethod evidence is essential. Without evidence to show that managerial abilities were in fact measured, program feedback may be misleading and detrimental. To use the Sackett and Dreher (1982) terminology, the multitrait-multimethod matrix findings can be either very troubling or not so troubling, depending on the stated purposes of the center.

Finally, it is also possible that the criterion-related validities would improve if abilities were better measured within exercises. Hypotheses pertaining to improved assessment center design should be empirically tested for this reason alone because, as Schmidt and Hunter would surely point out, the average criterion-related validity for the assessment center method does not compare favorably with that for cognitive ability tests (cf. Hunter & Hunter, 1984; Schmidt, Hunter, Pearlman, & Hirsh 1985).

References

- Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588–606.
- Borman, W. C. (1982). Validity of behavioral assessment for predicting military recruiter performance. *Journal of Applied Psychology*, 67, 3–9.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Endler, N. S., & Magnusson, D. (1976). Toward an interactional psychology of personality. *Psychological Bulletin*, 83, 956–974.
- Epstein, S. (1979). The stability of behavior: 1. On predicting most of the people much of the time. *Journal of Personality and Social Psychology*, 37, 1097–1126.
- Epstein, S. (1983). The stability of confusion: A reply to Mischel and Peake. *Psychological Review*, 90, 179–184.
- Epstein, S., & O'Brien, E. J. (1985). The person-situation debate in historical and current perspective. *Psychological Bulletin*, 98, 513–537.
- Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity [Monograph]. *Journal of Applied Psychology*, 72, 492–510.

- Harvey, R. J., Billings, R. S., & Nilan, K. J. (1985). Confirmatory factor analysis of the Job Diagnostic Survey: Good news and bad news. *Journal of Applied Psychology, 70*, 461-468.
- Hinrichs, J. R. (1978). An eight-year follow-up of a management assessment center. *Journal of Applied Psychology, 63*, 596-601.
- Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72-98.
- James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models, and data*. Beverly Hills, CA: Sage.
- Jöreskog, K. G., & Sörbom, D. (1983). *LISREL: Analysis of linear structural relationships by the method of maximum likelihood* (2nd ed.). Chicago: National Educational Resources.
- Kalleberg, A. L., & Kluegel, J. R. (1975). Analysis of the multitrait-multimethod matrix: Some limitations and an alternative. *Journal of Applied Psychology, 60*, 1-9.
- Kenrick, D. T., & Stringfield, D. O. (1980). Personality traits and the eye of the beholder: Crossing some traditional philosophical boundaries in the search for consistency in all of the people. *Psychological Review, 87*, 88-104.
- Long, J. S. (1983). *Confirmatory factor analysis*. Beverly Hills, CA: Sage.
- Mischel, W., & Peake, P. K. (1982). Beyond déjà vu in the search for cross-situational consistency. *Psychological Review, 89*, 730-755.
- Mischel, W., & Peake, P. K. (1983). Some facets of consistency: Replies to Epstein, Funder, and Bem. *Psychological Review, 90*, 394-402.
- Magnusson, D., & Endler, N. S. (1977). Interactional psychology: Present status and future prospects. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology* (pp. 3-31). Hillsdale, NJ: Erlbaum.
- Neidig, R. D., & Neidig, P. J. (1984). Multiple assessment center exercises and job relatedness. *Journal of Applied Psychology, 69*, 182-186.
- Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology, 67*, 401-410.
- Sackett, P. R., & Dreher, G. F. (1984). Situation specificity of behavior and assessment center validation strategies: A rejoinder to Neidig and Neidig. *Journal of Applied Psychology, 69*, 187-190.
- Schmidt, F. L., Hunter, J. E., Pearlman, K., & Hirsh, H. R. (1985). Forty questions about validity generalization and meta-analysis. *Personnel Psychology, 38*, 697-798.
- Schmitt, N., & Stults, D. M. (1986). Methodological review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement, 10*, 1-22.
- Shack, M. S. (1983). *The implications of cognitive and interpersonal processes for assessment centers*. Unpublished doctoral dissertation. Bowling Green State University.
- Silverman, W. H., Dalessio, A., Woods, S. B., Johnson, R. L. (1986). Influence of assessment center methods on assessors' ratings. *Personnel Psychology, 39*, 565-578.
- Statistical Analysis System. (1982). *SAS user's guide: Basics*. Cary, NC: SAS Institute.
- Sweeney, D. C. (1976). *The development and analysis of rating scales for the Chicago Police Recruit Assessment Center*. Unpublished manuscript, Bowling Green State University, Psychology Department.
- Thomson, H. A. (1970). Comparison of predictor and criterion judgments of managerial performance using the multitrait-multimethod approach. *Journal of Applied Psychology, 54*, 496-502.
- Turnage, J. J., & Muchinsky, P. M. (1982). Transsituational variability in human performance within assessment centers. *Organizational Behavior and Human Performance, 30*, 174-200.
- Turnage, J. J., & Muchinsky, P. M. (1984). A comparison of the predictive validity of assessment center evaluations versus traditional measures in forecasting supervisory job performance: Interpretive implications of criterion distortion for the assessment paradigm. *Journal of Applied Psychology, 69*, 595-602.
- Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*, 1-26.

Received September 7, 1983

Revision received February 2, 1987

Accepted December 8, 1986 ■

Estimating the Standard Error of Projected Dollar Gains in Utility Analysis

Ralph A. Alexander and Murray R. Barrick
University of Akron

Although attention given the utility analysis of personnel interventions has substantially increased recently, researchers have not addressed the problem of the standard error of utility estimates. The method for estimating such standard errors is presented and demonstrated in this article.

In the past few years there has been a resurgence of interest in utility or cost-benefit analyses of personnel interventions. Much of the impetus stems from Schmidt, Hunter, McKenzie, and Muldrow (1979). Their empirical method for estimating the dollar value of the standard deviation of performance (SD_y) brought about a renewed focus on utility research by providing one means for overcoming a serious obstacle to the application of utility estimation. Since 1979, alternative methods for estimating SD_y have been developed (Cascio, 1982; Cascio & Ramos, 1986; Eaton, Wing, & Mitchell, 1985) and a number of authors have called for wider use of utility analysis in personnel research (Cascio, 1980; Landy, Farr, & Jacobs, 1982; Schmidt, Hunter, & Pearlman, 1982). Although most utility analyses have evaluated selection procedures, recent work has extended its use to other human resource interventions such as recruitment (Boudreau & Rynes, 1985), training (Landy et al., 1982), promotions (Cascio & Ramos, 1986), and employee flow through the work force (Boudreau & Berger, 1985).

Throughout the literature on utility estimation there is an important omission. Neither in the early developmental work on utility analysis nor in the recent refinement, extension, and application literature do we find a consideration of the standard error of estimates of utility gains from personnel interventions. Although it is obviously useful for an organization to have an estimate of the expected value of returns from investments in human resources, the variance associated with such estimates is also crucial to organizational decision making, if for no other reason than to determine whether the confidence interval includes zero.

Most if not all of the variables entering a typical utility analysis have some variability (or uncertainty) associated with them (Alexander & Cronshaw, 1984; Alexander, Cronshaw, & Barrick, 1986; Cronshaw & Alexander, 1983, 1985), and most recent empirical studies of utility analysis report standard errors for at least some of these variables. In none of this literature is the extension made to the standard error of the overall utility estimate.

This article details the method for calculating this standard

error and demonstrates its use. Before proceeding with the calculations we review briefly both the classic utility model and recent capital budgeting modifications of that model.

Classic Utility Model

The classic utility model (Brogden, 1949; Cronbach & Gleser, 1965) is expressed as some variant of the following equation:

$$U = (T)(N_s)(SD_y)(r_{xy})(\lambda/\phi) - (C_a)(N_s)/(\phi), \quad (1)$$

where U = the dollar value payoff resulting from the human resource intervention (e.g., a selection program); T = the time period duration of the intervention (i.e., the average tenure of selectees); N_s = the number of individuals impacted (the number of selectees in the case of selection utility); SD_y = the standard deviation of performance in the present employee group in dollars; r_{xy} = the validity of the (selection) intervention as a predictor of performance; ϕ = the selection ratio; λ = the height of the normal curve associated with ϕ ; and C_a = the per applicant cost of implementing the selection program.

Cronshaw and Alexander (1983, 1985) and Boudreau (1983) modified this basic model to take the time value of money into account. Simply stated, this means that a dollar earned by an organization in the present year is worth more than a dollar earned at some future time, if for no other reason than that the present dollar can be invested at prevailing interest rates. Cronshaw and Alexander (1983, 1985) showed that this is easily accounted for by applying a capital budgeting model to Equation 1 to give the following equation:

$$U = \sum_{t=1}^T \frac{(N_s)(SD_y)(r_{xy})(\lambda/\phi)}{(1+i)^t} - (C_a)(N_s)/(\phi) \\ = \frac{(1+i)^T - 1}{i(1+i)^T} (N_s)(SD_y)(r_{xy})(\lambda/\phi) - (C_a)(N_s)/(\phi), \quad (2)$$

where i = the cost of capital (often referred to as the discount rate) and all other terms are as previously defined. In this form, U would be referred to as the net present value (NPV) of the intervention. Cronshaw and Alexander (1985) showed a similar derivation for other useful capital budgeting indexes such as internal rate of return, payback period, return on investment, and so on. A comparison of Equations 1 and 2 shows that they are

The authors would like to thank Steven F. Cronshaw and Frank L. Schmidt for their helpful comments on an earlier draft of this article.

Correspondence concerning this article should be addressed to Ralph A. Alexander, Department of Psychology, University of Akron, Akron, Ohio 44325.

identical except for the term associated with the time duration of the intervention: T in Equation 1 becomes $[(1 + i)^T - 1]/i(1 + i)^T$, which can be thought of as the *discount-adjusted* duration of the intervention. It is worth noting that to financial decision makers this capital budgeting modification is essentially a standardizing transformation that permits the comparison of projects that have different time frames. Discounting future returns is also a standard method for dealing with the fact that future forecasting becomes increasingly problematic for longer time spans (Beenhakker, 1976).

With these two forms of the utility estimate, we now turn to the problem of estimating the variance of U . The discussion will be facilitated by considering both Equations 1 and 2 in their generic form, with utility being the difference between returns and costs:

$$U = R - C. \quad (3)$$

Variance of a Function of Random Variables

To compute the variance in a composite function of variables, we begin by recalling that given some variable X and constants a and b such that $Y = a + bX$, the variance in Y is

$$S_y^2 = b^2 S_x^2. \quad (4)$$

Using this relation to find the variance in estimated U (from Equations 1 or 2), we must separate the variables (T , N_s , SD_y , etc. in the returns component, and C_a , N_s , and $1/\phi$ in the costs component) into those that are random variables or variables measured with error (the X s) and those that are considered constants. The product of the expected values of the constants will become b in Equation 4, and the variance of the product of the random variables will become S_x^2 . Both b and S_x^2 are found separately for the returns (R) and costs (C).

We are left, then, with finding an expression for the variance of the product of two or more random variables. The formula for the variance of the product of random variables when the population means and variances of the variables is known has been in the statistical literature since the 1930s and was (apparently) independently derived for use in meta-analysis in Hunter, Schmidt, and Jackson (1982, pp. 77, 85). Goodman (1960) dealt with the estimation of this variance in the two-variable case when the means and variances of the individual variables are sample estimates. Goodman (1962) generalized this to any number of random variables. His Equation 4, in which $X = \Pi X_i$, is the following:

$$S_x^2 = \prod [M_i^2 + S_i^2(n_i - 1)/n_i] - \prod (M_i^2 - S_i^2/n_i), \quad (5)$$

where Π is the product over the i random variables; S_i^2 is the variance of the i th variable; M_i^2 is the squared mean of the i th variable; and n_i is the sample size on which the mean and variance of the i th variable are based.

To summarize, then, we first find (separately) the variances in the cost component (S_C^2) and the returns component (S_R^2) of Equations 1 or 2 by the methods previously described. Because costs and returns can reasonably be expected to be independent of each other, the variance in the overall utility is simply the following equation:

$$S_U^2 = S_R^2 + S_C^2. \quad (6)$$

Application of the Method

Schmidt et al. (1979) analyzed the dollar-value utility that might be expected from using the Programmer Aptitude Test (PAT) to select computer programmers in the federal government. Their analysis was based on the classic utility model (Equation 1). Cronshaw and Alexander (1983) applied the capital budgeting modification of that model (Equation 2) to the Schmidt et al. (1979) data to demonstrate the NPV model calculations. Because the information needed to calculate the standard error of the utility estimates is available or can reasonably be estimated for the Schmidt et al. PAT utility analysis, that data will be used to demonstrate the method.

Schmidt et al. (1979) argued that utility analyses are often best conducted by analyzing the data for a range of values for the selection ratio. We will say more about this later; but for purposes of comparison, the utility analysis was conducted at four selection ratios (.10, .20, .50, and .80). For each of these, then, λ/ϕ is treated as a constant. The value of λ (the height of the normal curve) can be found for a particular selection ratio, ϕ , from tables of the standard normal distribution. The λ/ϕ constants for these selection ratios are 1.755, 1.40, 0.7978, and 0.350, respectively.

Schmidt et al. (1979) also reported values for SD_y ($\bar{M} = \$10,413$; $SD = 1,354$; $n = 105$), N_s ($\bar{M} = 610$; $SD = 45$; $n = 2$), an estimated true validity of .76, an average tenure (T) of 9.69 years, and a cost per applicant (C_a) of \$10. The r_{xy} estimate is from a validity generalization study, and Schmidt, Gast-Rosenberg, and Hunter (1980) reported an adjusted standard error of .25 ($n = 42$) for that value. (The determination of this n size will be discussed in detail later.) Schmidt et al. provided no information on the standard deviation or n size for their estimate of tenure. An informal survey of six large organizations, each of which employs more than 100 computer programmers, found a value of $SD_T = 3$ to be a reasonable approximation on the basis of an n of 325.

This provides all of the information needed to calculate the standard error of utility estimates, SE_U , under the classic model (Equation 1). For Equation 2, recall that T is replaced by a term reflecting the discount-adjusted time duration of the intervention. Most recent capital budgeting utility estimates have used a discount rate, i , of .10. The $[(1 + i)^T - 1]/i(1 + i)^T$ term for $i = .10$, and $T = 9.69$, is 6.029. The exact variance of that term is far too cumbersome for the present context. A sufficiently close approximation to the standard error can be obtained by multiplying the standard deviation in T by the ratio of the discount-adjusted time to the unadjusted time—that is $[(6.029/9.69) \times 3] = (.622 \times 3) = 1.867$. Simulations using values of T and i typical of utility analyses shows this approximation to underestimate the actual standard error by as much as 10%.

Table 1 gives a step-by-step example of the calculation of SE_U . Table 2 gives the values of U for selection ratios of .10, .20, .50, and .80 for both the classic and NPV utility models. The standard errors were found by first applying Equation 5 to find the variance in the product of four random variables (T , N_s , SD_y , r_{xy}), then Equation 4 was used with $b = \lambda/\phi$ to find the variance in the returns (S_R^2). Application of Equation 4 with the variance in N_s and $b = C_a/\phi$ yielded the variance in cost (S_C^2). Finally, S_U^2 was calculated as the sum of S_R^2 and S_C^2 (Equation 6).

Table 1
Sample Calculation of SE_U for the Classic Utility Model
With a Selection Ratio (ϕ) of .10

Equation
Step 1. Compute S_x^2 for the returns component (Equation 5)
$S_x^2 = \Pi[\bar{M}_i^2 + S_i^2(n_i - 1)/n_i] - \Pi(\bar{M}_i^2 - S_i^2/n_i);$
where
$X_1 = T; \bar{M}_1^2 = (9.69)^2; S_1^2 = (3.0)^2; n_1 = 325.$
$X_2 = N_s; \bar{M}_2^2 = (610)^2; S_2^2 = (45)^2; n_2 = 2.$
$X_3 = SD_y; \bar{M}_3^2 = (10413)^2; S_3^2 = (1354)^2; n_3 = 105.$
$X_4 = r_{xy}; \bar{M}_4^2 = (.76)^2; S_4^2 = (.25)^2; n_4 = 42.$
$S_x^2 = [(9.69)^2 + (3.0)^2 (324/325)][(610)^2 + (45)^2 (1/2)]$ $[(10,413)^2 + (1,354)^2 (104/105)][(.76)^2 + (.25)^2 (41/42)]$ $- [(9.69)^2 - (3.0)^2/325][(610)^2 - (45)^2/2]$ $[(10413)^2 - (1354)^2/105][(.76)^2 - (.25)^2/42].$
Step 2. Compute S_R^2 for the returns component (Equation 4)
$S_R^2 = (\lambda/\phi)^2 S_x^2 = (1.755)^2 S_x^2.$
Step 3. Compute S_C^2 for the costs component (Equation 4)
$S_C^2 = (C_a/\phi)^2 (S_{N_y}^2) = (10/.10)^2 (45)^2.$
Step 4. Compute SE_U (Equation 6)
$S_U^2 = S_R^2 + S_C^2$ $SE(U) = (S_U^2)^{1/2}.$

Discussion

The use of utility analysis to assess the potential dollar gains to be realized by an organization's personnel interventions is becoming an increasingly acceptable and valuable tool to the personnel psychologist. Just as in other areas of measurement, the variate values on which such analyses are based—and thus the overall utility estimates—are subject to uncertainty (measurement error). It is by reporting the standard error of a measure that we convey information regarding the relative magnitude of that uncertainty.

We emphasize that Equations 1 and 2, and, consequently, the standard error calculations presented here, apply to the instance in which the intervention is used on a single cohort (e.g.,

use of a new selection test for a single year's hirees). Cronshaw and Alexander (1983, 1985) and Boudreau (1983) discussed extensions of the capital budgeting model (Equation 2) to multiple cohorts. Extension of the methods presented here to the calculation of the standard error of utility for multiple cohorts is detailed in Alexander and Barrick (1986).

The capital budgeting literature uses three different approaches for evaluating the operating characteristics and boundary conditions of projected dollar gains from capital investments. The two most common approaches in the literature are sensitivity analysis and Monte Carlo simulations (Beenhakker, 1976). Sensitivity analysis is used to evaluate the relative contribution of each of the random variables to overall dollar gains, and simulations are used to evaluate the likelihood of specific values of expected gains. Cronshaw, Alexander, Wiesner, and Barrick (in press) demonstrated the application of both of these procedures to utility analysis. Statistical approaches similar to the approach in the present article are much less common in the capital budgeting literature (Wagle, 1967).

Inspection of Table 2 shows that for this particular data the standard errors of the utility estimates are relatively large. Financial decision makers may be more interested in the confidence interval of the projected gain from an investment than in the absolute value of the standard error. Table 2 also shows the 90% confidence interval of each estimate. In the present analysis we see that for the cases studied, the lower bound of the 90% confidence interval is well above zero, and thus the downside risk is minimal.

The analysis conducted here to demonstrate the method for calculating SE_U followed the lead of Schmidt et al. (1979) and treated λ/ϕ as a constant by analyzing utility at several selection ratios. Some situations, however, may call for the utility analysis to be performed for some empirically estimated selection ratio (Alexander, Hanges, Kollar, & Alliger, 1986), in which case it will be necessary to determine whether that value is best treated as a random variable or a fixed constant. If, for example, an organization uses a fixed cutoff on the predictor (Kroeck, Barrett, & Alexander, 1983), ϕ could be treated as a constant (Alexander, Barrett, & Doverspike, 1983). If, on the other hand, the selection ratio may vary, an estimate of the variance of λ/ϕ and of $1/\phi$ will also be needed. Alexander, Hanges, and Kollar (1986) tabled these values for empirically determined estimates of the selection ratio.

Table 2
Expected Value, Standard Error, and 90% Confidence Interval of Expected Dollar Gains From the Use of the Programmer Aptitude Test for One Year to Select Computer Programmers in the Federal Government^a

Selection ratio	Utility model					
	Classic (Equation 1) ^b			Net present value (Equation 2) ^c		
	U^a	SE_U	90% confidence interval	U^b	SE_U	90% confidence interval
.10	82.0	40.3	15.8–148.3	51.0	25.1	9.8–92.2
.20	65.5	32.1	12.6–118.3	40.7	20.0	7.8–73.6
.50	37.3	18.3	7.2–67.4	23.2	11.4	4.5–41.9
.80	16.4	8.0	3.2–29.6	10.2	5.0	2.0–18.4

^a Figures are in millions of dollars. ^b U s for the classic model are comparable to those reported by Schmidt, Hunter, McKenzie, and Muldrow (1979). ^c U s for the net present value model are comparable to those reported by Cronshaw and Alexander (1983).

The method is easily adapted to other situations. For example, in some settings the cost of implementing the intervention may be a constant, regardless of the number of individuals impacted (e.g., some training situations). In that case, C is a constant and referring back to the discussion of Equation 4, the variance in utility (S_U^2) will equal the variance in the returns (S_R^2). In a further refinement of the capital budgeting utility model (Equation 2), Boudreau (1983) included an additional multiplicative term in the returns (R) to account for organization costs that vary as a function of individual performance. He designated that term $1 + V$. Alexander, Cronshaw, Barrett, and DeSimone (1985) discussed the special case in which such variable costs may be estimated by a function of the pay-performance correlation. In these instances, an additional multiplicative term in R would be added to Equation 5.

This method for computation of the standard error requires two assumptions, and some discussion of them is in order. The first—required by Equation 5 for the variance of the product of random variables—is that the random variables, in this case T , N_s , SD_y , and r_{xy} , are uncorrelated with each other (Bohrnstedt & Goldberger, 1969). There is no reason to expect that SD_y would be correlated with any of the other variables. If the assumption is violated, the most likely possibility would be that r_{xy} correlates positively with T and negatively with N_s . That is, if the true validity were higher than the expected value used in the utility analysis, average tenure (T) might be greater and the required number of new hires (N_s) smaller than their sample-based expected values. It should be clear that such intercorrelations (in a particular single-cohort utility study) will be the result of correlated errors of estimation or correlated measurement errors. They will not arise if the usual assumptions of random and uncorrelated errors is met.

Although the statistical literature provides an adjustment to Equation 5 to deal with correlated variables (Bohrnstedt & Goldberger, 1969), such an adjustment is of limited usefulness because the information from which to estimate these intercorrelations will seldom be available. Note that the assumption of zero correlation among the random variables taken in this present application is the same assumption that is implicit in all utility analyses to appear in the industrial/organizational psychology (I/O) literature. Only if the variables are uncorrelated will the expected value of their product be equal to the product of their expected values. If the random variables are positively correlated, both the expected value (that is, Equations 1 and 2) and the standard error of the utility will be underestimated by the present methods (Goodman, 1960, 1962).

We emphasize that the assumption that the variables (T , N_s , etc.) are uncorrelated has to do with these intercorrelations in any single utility study. Intercorrelations among these variables across studies (jobs, organizations, etc.) are irrelevant to the assumption.

The second assumption is that the costs and returns are uncorrelated (Equation 6). The assumption is a reasonable one (Beenhakker, 1976), but once again the typical utility analysis setting will not be amenable to the estimation of the correlation. If costs and returns are positively correlated, Equation 6 will overestimate the standard error.

We comment on two other issues that arise from the use of this method. First, although the method is accurate for finding the standard error of U , use of that value for constructing accu-

rate confidence intervals depends on the distribution of U being approximately normal. In general, the distribution of U will be skewed (positively skewed when all expected values of the variables forming the product are positive; DeZur & Donahue, 1965). This skew will be at a maximum when the variables forming the product have equal variances and will become negligible as the ratio of the largest to second largest variance becomes quite large. As this ratio becomes very large, the distribution of the variable with the largest variance will come to dominate the shape of the distribution of U . (Olcott, 1973, demonstrated this in detail for the product of two variables.) In the usual utility analysis, the values encountered are likely to satisfy this requirement sufficiently well to alleviate the potential problem of a skewed U distribution. In the PAT data (Table 2), for example, SD_y has the largest variance, 1,354², and N_s the next largest, 45². The ratio of these two values is extremely large. Because for most utility analyses of personnel interventions, SD_y will likely have the largest variance, the shape of the distribution of SD_y will dominate the shape of the distribution of U . Schmidt et al. (1979) concluded that the distribution of SD_y for the PAT was not significantly nonnormal. In applications in which there is evidence for a substantial skew in SD_y , the upper and lower confidence bounds for U should be calculated separately using the separate standard errors for SD_y from the upper and lower halves of the distribution.

The second issue is the matter of n sizes. In most cases the n sizes for Equation 5 will be straightforward. The demonstration in this article, however, illustrates where a problem may arise. Recall that the value for r_{xy} and its standard error were from a validity generalization study (Schmidt et al., 1980). The adjusted standard error of r_{xy} in that study was a function both of the number of studies ($N = 42$) and the total number of subjects ($N = 1,299$) in the meta-analysis. Statistical theory provides no guidance as to the "proper" n size for Equation 5 in such instances. The conservative approach is to use the smaller n , which was done here.

Alexander, Cronshaw, and Barrick (1986) and Cronshaw and Alexander (1985) emphasized that in order for personnel interventions, such as selection systems, training programs, and so forth, to be treated on an equal footing with other investment decisions by organizations, it is necessary to abandon a cost mentality and treat the costs of obtaining, improving, and retaining valued human resources as investments that are expected to generate returns to the organization over an extended period of time. This shift of focus requires both a change in certain of the features of the utility model and in expressing the results of such analyses in a language that is well understood by financial decision makers. The classic Cronbach-Gleser utility model ignores a number of factors that are routinely dealt with in organizational financial analysis. Such considerations lead Cronshaw and Alexander (1983, 1985) and Boudreau (1983) to recast the classic utility model into the capital budgeting framework of Equation 2. Reference to Table 2 shows that this model has two other advantages over the classic model. The first is that it produces more conservative (and in the view of the financial decision makers, more comparable) estimates of utility. The NPV model also has a substantially smaller standard error.

The method is now available for calculating the standard error of estimated utility gains from personnel interventions. It is recommended that future studies report not only the expected

value of such gains but also the standard error of these estimates.

References

- Alexander, R. A., Barrett, G. V., & Doverspike, D. (1983). An explication of the selection ratio and its relationship to hiring rate. *Journal of Applied Psychology*, 68, 342-344.
- Alexander, R. A., & Barrick, M. R. (1986). *Estimating standard errors for utility analysis with multiple cohorts*. Manuscript submitted for publication.
- Alexander, R. A., & Cronshaw, S. F. (1984, August). *The utility of selection programs: A finance-based perspective*. Paper presented at the 92nd Annual Convention of the American Psychological Association, Toronto, Canada.
- Alexander, R. A., Cronshaw, S. F., Barrett, G. V., & DeSimone, R. L. (1985). *Productivity gains and organizational benefits: Accounting for the pay-performance relationship in utility analysis*. Manuscript submitted for publication.
- Alexander, R. A., Cronshaw, S. F., & Barrick, M. R. (1986, April). *Extending the managerial finance model of utility analysis to deal with uncertainty in parameter estimates*. Paper presented at the 1st Annual Conference of the Society for Industrial and Organizational Psychology, Chicago.
- Alexander, R. A., Hanges, P. J., & Kollar, L. (1986). *Standard errors of sample-based estimates of the selection ratio*. Manuscript submitted for publication.
- Alexander, R. A., Hanges, P. J., Kollar, L., & Alliger, G. M. (1986). *Estimating parameters under range restriction when the population variances are unknown*. Manuscript submitted for publication.
- Beenhakker, H. L. (1976). *Handbook for the analysis of capital investments*. Westport, CT: Greenwood.
- Bohrnstedt, G. W., & Goldberger, A. S. (1969). On the exact covariance of products of random variables. *Journal of the American Statistical Association*, 64, 1439-1442.
- Boudreau, J. W. (1983). Effects of employee flows on utility analysis of human resource productivity improvement programs. *Journal of Applied Psychology*, 68, 396-406.
- Boudreau, J. W., & Berger, C. J. (1985). Decision-theoretic utility analysis applied to employee separations and acquisitions [Monograph]. *Journal of Applied Psychology*, 70, 581-612.
- Boudreau, J. W., & Rynes, S. L. (1985). Role of recruitment in staffing utility analysis. *Journal of Applied Psychology*, 70, 354-366.
- Brogden, H. E. (1949). When testing pays off. *Personnel Psychology*, 2, 171-183.
- Cascio, W. F. (1980). Responding to the demand for accountability: A critical analysis of three utility models. *Organizational Behavior and Human Performance*, 25, 32-45.
- Cascio, W. F. (1982). *Costing human resources: The financial impact of behavior in organizations*. Boston: Kent.
- Cascio, W. F., & Ramos, R. A. (1986). Development and application of a new method for assessing job performance in behavioral/economic terms. *Journal of Applied Psychology*, 71, 20-28.
- Cronbach, L. J., & Gleser, G. (1965). *Psychological tests and personal decisions*. Urbana: University of Illinois Press.
- Cronshaw, S. F., & Alexander, R. A. (1983). The selection utility model as an investment decision: The greening of selection utility. *Proceedings of the 43rd Annual Meeting of the Academy of Management*, 297-300.
- Cronshaw, S. F., & Alexander, R. A. (1985). One answer to the demand for accountability: Selection utility as an investment decision. *Organizational Behavior and Human Decision Processes*, 35, 102-118.
- Cronshaw, S. F., Alexander, R. A., Wiesner, W. H., & Barrick, M. R. (in press). Incorporating risk into selection utility: Two models for sensitivity analysis and risk simulation. *Organizational Behavior and Human Decision Processes*.
- DeZur, R. S., & Donahue, H. D. (1965). *On the products and quotients of random variables* (Report No. ARL 65-71). Washington, DC: Aerospace Research Laboratory.
- Eaton, N. K., Wing, H., & Mitchell, K. J. (1985). Alternate methods of estimating the dollar value of performance. *Personnel Psychology*, 38, 27-40.
- Goodman, L. A. (1960). On the exact variance of products. *Journal of the American Statistical Association*, 55, 708-713.
- Goodman, L. A. (1962). The variance of the product of K random variables. *Journal of the American Statistical Association*, 57, 54-60.
- Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis*. Beverly Hills, CA: Sage.
- Kroeck, K. G., Barrett, G. V., & Alexander, R. A. (1983). Imposed quotas and personnel selection: A computer simulation study. *Journal of Applied Psychology*, 68, 123-136.
- Landy, F. J., Farr, J. L., & Jacobs, R. R. (1982). Utility concepts in performance measurement. *Organizational Behavior and Human Performance*, 30, 15-40.
- Olcott, R. J. (1973). Confidence limits on the product of two uncertain numbers. *Analytical Chemistry*, 45, 1737-1740.
- Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J. E. (1980). Validity generalization results for computer programmers. *Journal of Applied Psychology*, 65, 643-661.
- Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). Impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology*, 64, 609-626.
- Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1982). Assessing the economic impact of personnel programs on work-force productivity. *Personnel Psychology*, 35, 333-347.
- Wagle, B. (1967). A statistical analysis of risk in capital investment projects. *Operational Research Quarterly*, 18, 13-32.

Received February 10, 1986

Revision received November 26, 1986

Accepted January 12, 1987 ■

Use of Tests Manifesting Sex Differences as Measures of Intelligence: Implications for Measurement Bias

Mary Roznowski
University of Illinois, Urbana-Champaign

Very narrow tests measuring knowledge of specific information from Project TALENT were combined into two composites on the basis of between-group differences for high-school-age boys and girls. These composites were analyzed to determine what happens when specific, nontrait components of variance are included in measures of general intelligence. The two composites were heavily advantageous to either males or females and were made up of very narrow, mostly nonacademic, information-dependent subtests. Correlations were computed between the sex-advantage composites and general intelligence scores. Very large validities were obtained, indicating that the composites were acting as excellent measures of general intelligence for both sexes. Results are discussed in the framework of multiple determinants of responses and group differences in item and test performance.

Any one item or even a narrow subtest is an extremely fallible measure of intelligence; neither items nor tests can be pure measures of any meaningful hypothetical trait. Rather, item responses are determined by multiple factors including, of course, the trait of interest. Experiences associated with such factors as sex, socioeconomic status, race, religion, and ethnicity, among others, may be determinants or correlates of responses to intelligence test items. However, the presence of such response determinants does not necessitate removing an item if variance in item responses is substantially related to the trait being measured. Problems arise only when responses to items in a test are determined in large part by irrelevant and inappropriate factors and not by the construct of interest.

Humphreys has long argued against high homogeneity on the basis of both predictive and construct validity concerns (Humphreys, 1952, 1962, 1970, 1981, 1986). Although seemingly paradoxical, he points out that the only reasonable way to keep nontrait determinants from contributing significantly to test scores is to deliberately include a diversity of items with heterogeneous determinants (both attribute and systematic nontrait determinants) in test batteries. With a fixed number of items, the larger the number of different, nontrait determinants, the less any individual determinant will contribute to total test variance. Eliminating items from a test because their content capitalizes on experiences of, or knowledge more likely attained by, certain subgroups or cultures may impair the test's predictive

power. This is because the test essentially becomes too homogeneous as contributions from still other nontrait determinants in remaining items increase.

In constructing a test, it is best to combine items that are related to the trait of interest and, in addition, have a diversity of correlates, antecedents, and determinants (the term *trait-relevant* will be used to signify this idea). This diversity should increase the contributions from attribute or trait sources and decrease the contribution to total test variance from each systematic, nontrait source and ultimately improve the validity of the test.

This is a paradoxical situation. When such systematic heterogeneity is reduced by including a set of items with fewer nontrait determinants, each individual, nontrait determinant remaining in the item pool will add relatively more to total test variance, because items sharing one or more of the determinants will covary with each other and consequently contribute more to item-item covariances. Except for very short tests, the $N(N - 1)$ covariance terms far outweigh the N variance terms in determining total test variance. If items are selected to maximize trait-relevant heterogeneity, the covariance of any one item with the other items will be determined more likely by the attribute components rather than by any specific nontrait component.

An investigator can effectively deal with systematic, nontrait variance through a broad sampling of items from the available universe of trait-relevant measures. Questionable sources of variance can never be eliminated entirely, but they can be controlled by systematically maximizing heterogeneous, trait-relevant components.

The purpose of this study was to examine the effects of deliberately including items (actually, entire subtests were used) containing systematic, nontrait variance. In order to do this, subtests that contained both attribute variance and systematic, nontrait variance from a wide variety of sources were combined into composites. Individual subtests were chosen for composites if they were very narrow in content and favored either males

The author greatly appreciates the many helpful comments and suggestions offered by Lloyd Humphreys, Charles Hulin, and Timothy Davey. Professor Humphreys also made available the Project TALENT data set that he obtained from the Project TALENT Data Bank. For a description of the original project and data base, see Flanagan et al. (1962). Project TALENT's generosity in sharing their extensive data bank made this research possible.

Correspondence concerning this article should be addressed to Mary Roznowski, Department of Psychology, 603 East Daniel Street, Champaign, Illinois 61820.

or females. This latter condition was intended to ensure adequate nontrait variance components in the composites. A rural-urban demographic distinction could have been chosen as well. Furthermore, a very wide range of these narrow tests that differentiated the two groups was necessary. This requirement would increase heterogeneity in the resulting composites.

The subtests chosen ranged from knowledge of cooking to knowledge of the Bible to knowledge of foreign travel. The various systematic components in these subtests were all subgroup-relevant because they differentiated among individuals on the basis of sex. (No direct causal statements can be made, of course.) Because many very diverse subtests were chosen, the tests shared little, other than attribute and sex variance. For instance, the specific experiences associated with Bible knowledge are unlikely to be the same as the experiences associated with cooking knowledge. Determinants of farming knowledge likely differ from determinants of knowledge about foreign travel. Thus, the condition of a diversity of components with heterogeneous determinants and correlates seemed to be fulfilled. The requirement of large sex differences served to provide a systematic criterion with which to choose tests.

A criterion of general intelligence was used. It was predicted that correlations between general intelligence and the male and female advantage composites (the composites made up of tests favoring males or females) would be low to moderate for the cross-sex analysis (i.e., male advantage test as a predictor of intelligence in the female sample) if the composites were providing poor measurement of individual differences in intelligence. More generally, poor correlations with intelligence were expected if the narrow, information-dependent subtests acted as little more than noise or "bias" factors contributing little to the prediction of general intelligence. Sizable correlations between advantage composites and the criterion would indicate support for the idea that systematic heterogeneity does not degrade the measurement of the broad construct of general intelligence; the correlations would suggest that composites made up of tests influenced by "bias" factors possessed substantial validity for an unbiased criterion.

Method

The sample for this study was split into two subsamples to determine if the composites performed differentially in male and female samples. The two samples were composed of 10th-grade boys and girls tested in the 1960 Project TALENT nationwide testing. Subgroup *n*s for the samples were approximately 12,690 boys and 13,350 girls. The samples were representative of the population of 10th-grade students.

Table 1 contains the subtests chosen for each sex-advantage composite. Two composites were created, each having 20 components. If there was a substantial difference in means between males and females and the test was an assessment of very specific information knowledge, the test was included in the composite. For the majority of the tests, differences of at least one standard deviation between the groups were observed. The two composite tests were formed as unit-weighted, linear combinations of the subtests.

The measure of general intelligence chosen as the criterion measure was the Project TALENT intelligence composite that contained reading comprehension, arithmetic reasoning, and abstract reasoning. These are all found in standard tests of intelligence. Trivial mean differences were found between the two groups on this composite: The mean for the male sample was 163.95 and the mean for the female sample was

Table 1
Subtests Chosen for Each Sex-Advantage Composite

Female advantage	Male advantage
Music	Physical science
Home economics	Biological science
Art	Aeronautical and Space
Health	Electricity and Electronics
Clerical	Mechanics
Bible	Farming
Colors	Sports
Etiquette	Law
Theatre and Ballet	Engineering
Foods	Foreign travel
Arithmetic computation	Military
Table reading*	Hunting
Clerical checking*	Fishing
Object inspection*	Outdoor activities
Disguised words	Mechanical reasoning
Spelling	Creativity
Punctuation	2nd visualization
English usage	3rd visualization
Effective expression	Vocabulary I
Word functions in sentences	Introductory math

Note. Subtests are taken from the 1960 Project TALENT nationwide testing battery. Asterisk (*) indicates highly speeded tests.

161.31. With a standard deviation across both groups of 50.77, this difference of 2.64 represented .05 standard units.

Correlations were computed between the advantage composites and intelligence scores separately in each sample. In order to estimate reliability of the two composites, correlations between odd and even halves of each composite were computed in each sample. The Spearman-Brown formula was used to extend this split-half correlation to an estimate of reliability of the entire composite. Validities computed between sex-advantage composites and intelligence scores were corrected for attenuation using the reliability estimates.

Next, correlations between intelligence scores and the combined advantage composites (male and female composites, equally weighted) were computed in both samples. A correction for attenuation due to unreliability was applied to these validity coefficients. The correlation between male and female composites (considered as halves of the combined information composite) was used as a lower-bound estimate of reliability of the combined composites. Indeed, these halves cannot be considered to result from a random split of the components but from a rather "biased" split of the components. The resulting reliability values were used to correct the correlations between intelligence and combined superiority.

Results

Within-sex (e.g., male test in male sample) and cross-sex (e.g., female test in male sample) correlations between advantage composites and intelligence, standard deviations of advantage composites, and composite means are shown in Table 2. Reliabilities estimated from an application of the Spearman-Brown formula to the correlations between odd and even halves of the advantage composites are also given. Finally, validity correlations corrected for unreliability in the advantage composite are presented. The cross-sex validities have been highlighted to underscore the importance of these correlations.

The results in Table 2 show that heterogeneous tests contain-

Table 2
Correlations Between Advantage Composites and Intelligence, Means and Standard Deviations and Reliabilities for Composites, and Correlations Corrected for Unreliability in Composites

Composite	<i>r</i>	<i>M</i>	<i>SD</i>	<i>r'</i> _{xx}	<i>r</i> _{T_xy}
Male test in male sample	.81	139.51	20.30	.93	.84
Male test in female sample	.82	110.47	18.04	.92	.86
Female test in male sample	.71	181.14	19.74	.88	.76
Female test in female sample	.75	211.59	20.34	.88	.80

Note. The cross-sex validities are in boldface in order to underscore their importance relative to the same-sex validities.

ing many diverse but trait-relevant components, each distinctly favoring one subgroup, fare very well as measures of intelligence. When accumulated into broad composites, these very narrow, information-dependent tests very closely approximated tests of general intelligence for individuals *regardless of group membership*. The male-advantage composite was an especially effective measure of general intelligence. The female-advantage composite was somewhat less highly related to intelligence than the male-advantage test in both groups. An inspection of the tests chosen for the female composite revealed that three highly speeded “clerical” tests were included in the composite. These tests—although differentiating females from males—contained very little attribute variance as determined by within-sex correlations between each test and intelligence. Clearly, these three tests that were not highly trait-relevant would have been eliminated from further use in any conventional analysis because of their lack of discrimination. After eliminating these three tests, the correlations between the female composite and intelligence increased to those of the male composite despite a smaller number of components ($r_{\text{male}} = .81$, $r_{\text{female}} = .83$).

Correlations between the two advantage composites were moderately high but clearly not unity ($r_{\text{males}} = .69$, $r_{\text{females}} = .75$). Indeed, the composites overlapped in terms of sharing common attribute variance, although not to as large an extent as each did with the measures of intelligence. The two advantage composites were then combined into one overall composite to determine if the sum of the two composites would balance the sex differences and result in a higher correlation with the criterion than either composite alone. Obtained correlations were slightly higher ($r_{\text{males}} = .83$, $r_{\text{females}} = .84$). After correcting for unreliability in this overall composite, very large correlations were obtained ($r_{\text{males}} = .90$, $r_{\text{females}} = .91$). These correlations indicate that the overall composite was acting as a very effective measure of intelligence, even though it was constructed using elements containing specific, nontrait components of variance.

The stress of this article is on the theoretical importance of the correlations as estimates of validity. The predictive benefits gained from using the advantage composites are not of major importance. Furthermore, slopes and intercepts are not invariant parameters of regression equations. With a change of criterion from the intelligence test used here to a different measure of intelligence, one would likely derive different values for slopes and intercepts. A slight change in the composition of the advan-

tage composites would also likely have the effect of changing these regression estimates. Nonetheless, the regressions of the intelligence scores on the combined, sex-advantage composites in the two samples were computed for applied prediction interests. These regression equations were $\hat{Y} = 1.21X - 224.68$ (for the male sample) and $\hat{Y} = 1.23X - 235.47$ (for the female sample), where \hat{Y} is the intelligence score and X is the score on the combined, sex-advantage composites. This equivalence of the regression equations across samples confirms the usefulness of these sex-advantage composites as unbiased predictors of intelligence. It is obvious that no adjustment in the prediction equations is needed to achieve adequate and unbiased prediction in either sample. The interpretation of the theoretical correlations is thereby strengthened.

Discussion

Increasingly, items with systematic determinants that are not believed to be central to the underlying trait of interest have been considered undesirable and removed from tests. Objecting to test items because of group differences in performance is not a trend restricted to scientific researchers. In a recent legal case, the criterion for item bias was difficulty differences between groups (*Golden Rule Insurance v. Washburn*, 1984). This article is directly relevant to this decision. Items manifesting group differences are *not* necessarily biased, nor do poor measures of intelligence result from the use of such items.

Drasgow (1987) presented empirical results that converged with the results of this study. Although starting from a substantially different theoretical perspective, Drasgow found that the presence of biased items, identified by item-response theory methods, did not result in biased scale scores in the American College Testing (ACT) Assessment. The present study has shown that items (subtests here) with substantial group differences do not necessarily degrade measurement characteristics of the test as a whole. The two sets of results suggest that a different perspective on bias at the level of individual items is necessary.

Test composites containing systematic, nontrait components of variance actually result in very good relations with a theoretically and practically important criterion. The sex-advantage composites—narrow, information-dependent subtests—showed substantial within-group discrimination, so that intelligent students (regardless of group membership) performed well. The criticism that the criterion measure was assessing similar “bias” content as were the superiority tests is invalid; the composites resulted in substantial between-group differences, but the criterion test did not.

Sex differences—as well as many other subgroup differences—in item and test performance are ubiquitous. Sex differences exist on individual items in tests as well as in test scores from one narrow information test to another, as was found in this study and also noted by Humphreys (1986). Furthermore, any well-constructed test contains items whose responses are related to such factors as an examinee’s race, sex, ethnicity, religion, and to myriad psychological or other demographic or defining features of individuals. These differences will occur in both directions, and their existence does not necessarily mean the items are biased. As found in this study, a total

score made up of male and female advantage composites yielded high validity when the criterion was general ability. This occurred because a diverse and highly trait-relevant set of subparts (items or subtests) made up the composites.

A primary use of tests of cognitive abilities is the prediction of performance on some criterion. If one's goal is accuracy of prediction, eliminating items because of systematic, nontrait components may hinder achieving that goal. Indeed, contributions to variance, regarded as "bias" factors or components, are unavoidable and will not adversely affect the validity of the test if the components are sufficiently diverse and spread throughout the items. Test "purification" by removing the offending items may only contribute to a highly homogeneous item set that would very likely decrease validity and predictive accuracy and, paradoxically, increase the contribution to total test variance of other nontrait determinants. Psychometric fine-tuning may not achieve the ultimate goals of high validities for predicting important criteria and accurate measurement of the underlying trait or construct.

It must be pointed out that combining heterogeneous items or subtests that manifest group differences will not in itself accomplish the goals of high predictive and construct validities for a test. It is fundamental that *trait-relevant* items or subtests make up the test. An example is a test of knowledge of Black slang, which is relevant for correctly answering items written in Black urban argot. (The Black Intelligence Test of Cultural Homogeneity [BITCH] is one test containing such items [Williams, 1975]). Knowledge of such terms is largely irrelevant for individual differences in intelligence or even verbal ability. (Correlational analysis between BITCH and WAIS scores typically reveal very modest relations while, nonetheless, showing substantial group differences [cf. Matarazzo & Wiens, 1977]).

Many psychological measures have the natural and unavoidable consequence of being related to demographic and other nontrait variables. Between-group differences on an item or scale indicate very little about the relevance of within-group differences on the same items or scales for the measurement

of intelligence or other, more specific abilities. The strategy of selecting broadly within the theoretical limits of the domain of a trait is an excellent method for test constructors and users concerned about systematic, unwanted determinants and correlates of variance in test scores.

References

- Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, 72, 19-29.
- Flanagan, J. C., Darley, J. T., Shaycoft, M. F., Gorham, W. A., Orr, D. B., & Goldberg, I. (1962). *Design for a study of American youth*. Boston: Houghton Mifflin.
- Golden Rule Insurance Company et al. v. Washburn et al., No. 419-76 (stipulation for dismissal and order dismissing cause; Circuit Court of the Seventh Judicial Circuit, Sangamon County, IL, 1984).
- Humphreys, L. G. (1952). Individual differences. In C. P. Stone & D. Taylor (Eds.), *Annual Review of Psychology*, 3, 131-147.
- Humphreys, L. G. (1962). The organization of human abilities. *American Psychologist*, 17, 475-483.
- Humphreys, L. G. (1970). A skeptical look at the factor pure test. In C. E. Lunneborg (Ed.), *Current problems and techniques in multivariate psychology: Proceedings of a conference honoring Professor Paul Horst* (pp. 23-32). Seattle: University of Washington.
- Humphreys, L. G. (1981). The primary mental ability. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.), *Intelligence and Learning* (pp. 87-102). New York: Plenum Press.
- Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Journal of Applied Psychology*, 71, 327-333.
- Matarazzo, J. D. & Wiens, A. N. (1977). Black Intelligence Test of Cultural Homogeneity and Wechsler Adult Intelligence Scale scores of Black and White police applicants. *Journal of Applied Psychology*, 62, 57-63.
- Williams, R. L. (1975) *Black Intelligence Test of Cultural Homogeneity: Manual of directions*. St. Louis, MO: R. L. Williams & Associates.

Received March 17, 1986

Revision received February 16, 1987

Accepted January 26, 1987 ■

SHORT NOTES

Reliability and Validity of the Situational Interview for a Sales Position

Jeff A. Weekley and Joseph A. Gier
Zale Corporation, Irving, Texas

The reliability and validity of a situational interview (e.g., Latham, Saari, Pursell, & Campion, 1980) was examined for a sales position. In a pilot study, the interrater reliability of the interview was higher than that typically observed for interviews ($r = .84$). However, the estimate was probably inflated because the reliability was computed on the same data used in the item analysis phase of interview development. In a predictive validation study, the situational interview was shown to be valid ($r = .45$) in the prediction of sales productivity. After correcting for attenuation in the criterion, a validity coefficient of .47 was observed. Future research directions with respect to the situational interview are discussed.

The questionable reliability and validity of the employment interview has not prevented it from becoming the most widely used selection tool in use today (Arvey & Campion, 1982; Schmitt, 1976; Ulrich & Trumbo, 1965; Wagner, 1949). Because managers are apparently committed to the use of selection interviews, it is important that research continue to focus on ways to improve the reliability and validity of the procedure. The most promising avenues for future research and application are the various behaviorally based approaches to selection interviewing that are beginning to gain attention (e.g., Janz, 1982; Latham & Saari, 1984; Latham, Saari, Pursell, & Campion, 1980; Orpen, 1985).

One of these approaches is based on the well-founded premise that past behaviors will predict future behaviors (Ghiselli, 1966; Janz, 1982; Orpen, 1985). In general, the emphasis of this approach is on evaluating the candidate's behavior in past situations and generalizing to likely performance in future situations. Applicants who have performed successfully on past jobs are expected to be successful performers on the new job. Obviously, the greater the similarity between past work activities and those required by the new job, the more confident the interviewer may be in his or her generalizations. Using a variation of this technique, called patterned behavior description interviews (PBDI), Janz (1982) found lower reliability (.46 vs. .71) but higher validity (.54 vs. .07) for this procedure over the standard interview. Orpen (1985) subsequently replicated Janz' (1982) findings with respect to the superior validity of the PBDI over traditional interviewing procedures (.61 vs. .05).

Although the PBDI shows significant promise as a means of increasing interview validity, it suffers one potential limitation. If the applicant has never been confronted with tasks similar to those found in the new job, or if he or she has otherwise not

had the opportunity to engage in key behaviors in the past, the usefulness of the procedure may be reduced. Another approach, called the situational interview (Latham & Saari, 1984; Latham et al., 1980), seemingly overcomes this potential drawback. The situational interview is based on the premise that a person's expressed behavioral intentions are related to subsequent behavior. Thus, in the situational interview applicants are asked to describe how they think they would respond in certain job-related situations. In a series of concurrent and predictive validation studies, Latham, Saari, and colleagues (Latham & Saari, 1984; Latham, et al., 1980) have shown the validity of the situational interview to range from .14 to .46.

Our research on the situational interview was conducted in response to the sponsoring organization's need for a standard, job-related interviewing procedure for use in its 1,300 retail outlets. The research was conducted on a different type of job than that examined in the Latham and Saari studies (professional sales position) and used an objective performance measure (sales volume) as a criterion. The results of this study are reported in the interest of adding to the literature an independent assessment of the reliability and validity of the situational interview.

Method

Development of the Situational Interview

A situational interview form was developed following the method outlined by Latham et al. (1980). The first step involved the collection of critical incidents for the focal position, that of a sales associate in a jewelry retail organization. The incidents were gathered via a mail-out questionnaire containing a description of a complete critical incident (including what led up to the incident, what the sales associate did that was so effective or ineffective, and what were the consequences of his or her actions) as well as descriptions of two actual incidents (one positive and one negative). The questionnaire, which was sent to a sample of sales associates, assistant managers, and store managers, asked the respondents to relate six critical incidents for the sales associate position. This effort produced approximately 400 usable incidents of sales associate performance.

We would like to thank Blake Frank for his comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Jeff A. Weekley, MS 5B-2, P. O. Box 152777, Irving, Texas 75015-2777

Table 1
*Example Item From a Situational Interview
 for the Position of Sales Associate*

A customer comes into the store to pick up a watch he had left for repair. The repair was supposed to have been completed a week ago, but the watch is not back yet from the repair shop. The customer becomes very angry. How would you handle this situation?

- 1 ■ Tell the customer it isn't back yet and ask him or her to check back with you later.
- 2 ●
- 3 ■ Apologize, tell the customer that you will check into the problem and call him or her back later.
- 4 ●
- 5 ■ Put the customer at ease and call the repair shop while the customer waits.

These incidents were examined and provided the source from which a pool of 36 questions was written in a situational format. Specifically, most of the items were written so that a description of a situation was followed by the question "How would you handle that situation?" or "What would you have done in this situation?". The majority of the items dealt with basic sales and customer-service situations, whereas a few focused on security and prioritizing sales opportunities over non-sales tasks. The next step was to develop, for each question, a rating scale in which the 1, 3, and 5 scale points were anchored with answers that exemplified *below average*, *average*, and *above average performance*, respectively. To do this, a group of job experts was convened to review the questions. The job experts consisted of two members of the corporate training department (who were responsible for the development of sales skills training for the company) and two middle-level managers of store operations. The job experts, together with the first author, reviewed the questions and attempted to generate anchors for the three scale points. Discussion ensued concerning each suggested anchor until the four experts came to agreement regarding the best anchor for the scale point in question. For 6 of the questions it proved extremely difficult to come to a consensus regarding one or more of the anchors. These items were deleted from further consideration. Consensus was reached regarding appropriate anchors for the other 30 questions. An example of 1 situational question and its corresponding anchors can be found in Table 1.

Pilot Study

The next step in the construction of the final situational interview form was to conduct a pilot test of the preliminary items. In order to accomplish this, 14 store managers were contacted and agreed to serve as "interviewers." (Store managers were chosen because they were ultimately to be the users of the interview.) Twenty-one employees in the corporate headquarters were contacted and agreed to serve as "interviewees." In each case, the interviewee had no previous experience working for the company in a sales capacity and, therefore, should have been a reasonable proxy for the typical applicant for a sales associate job. Although most of the interviewee sample were nonexempt employees, a few were exempt-level professionals. In conducting the pilot inter-

views, the store managers were briefly trained in the use of the situational interview and then paired off. Each pair of store managers was to interview three of the applicants. The store managers took turns as the interviewer for an entire interview (as opposed to trading off questions). While one manager was interviewing, the other was instructed to sit quietly and unobtrusively off to the side and make his or her ratings of the interviewee's answers.

This process resulted in 42 sets of interview ratings, 2 for each of the interviewees. Unfortunately, one pair of store managers misunderstood the instructions and read the anchors as well as the questions to the interviewees. The 6 interview ratings from this pair were excluded from further analyses. Additionally, data for one of the interviewees had to be discarded (the managers indicated that she did not take the activity seriously and began giving answers that she considered funny). Thus, the pilot consisted of 17 interviews conducted by six pairs of store managers. After the interviews were complete, the store managers were debriefed. The debriefing focused on identifying ambiguous questions, questions that the managers thought would require knowledge of company policy to answer correctly, anchors that the managers thought were incorrect or ambiguous, and so forth. As a result of the debriefing, two of the items were deleted and two others slightly reworded.

From the pilot data, an item analysis was conducted. In the interest of producing a short, easily interpreted interview form (interview time was scarce in the actual store environment), an a priori decision was made to limit the final form to approximately 15 questions that could be summed for scoring. Thus, the goal of the item analysis was to identify 15 questions that maximized both interrater reliability and the total score variance. Because of the small sample size, item means and variances were computed across the 34 interview ratings available (the observations, of course, were not independent), whereas item reliabilities were computed across the 17 pairs of interview ratings. After eliminating 6 items having relatively low variances or reliabilities, the remaining items were examined and selected for inclusion on the basis of item content (the objective being to sample as broadly as possible from the range of duties represented in the situational questions). The final result was a 16-item interview form with an interrater reliability of .84 (note that this estimate is likely to be inflated, given the potential for capitalization on chance in the item analysis procedure). The internal consistency of the form was .61 (Cronbach's alpha computed across the 34 sets of ratings), which is not surprising given the intentional heterogeneity in item content.

Subjects and Procedure

The subjects in the validation phase of the study were 54 applicants for the position of sales associate. All of the subjects had previously worked for a major department store in the fine jewelry department. At the time of this study, the jewelry departments in these stores had been taken over by an outside organization that leased the jewelry counter space from the host department store. As part of the lease agreement, the outside organization was obligated to consider for employment any sales associate working in one of the host stores' fine jewelry departments at the time of the agreement. In practice this generally meant that, of the associates it "inherited," the outside organization had to hire all who wanted a job.

The operations manager responsible for assessing the applicants was contacted and asked to use the situational interview form in his interviews. He agreed to use the interview and to forward the completed forms, along with the person's name and social security number, to headquarters for subsequent analysis. The division personnel manager, who had previously been given training in the use of the situational interview, trained the operations manager on site in the proper use of the procedure. All interviews were conducted by either the operations manager or the personnel manager.

Of the 54 candidates interviewed, 24 were subsequently hired as sales associates. Of those interviewed but not hired, some were not offered jobs (e.g., because they were deemed a high security risk), some were offered jobs but turned them down, and others were absorbed into the host stores' other departments. In order to examine the possibility that the situational interview was used to select from the original sample of 54 candidates, the total interview score of the hired group ($N = 24$) was compared with that of the nonhired group ($N = 30$). Results of the t -test showed that there was no difference between the overall interview score for the two groups, substantiating the assertion that the lease agreement effectively precluded actual selection. This also suggests that range restriction, created by selecting on the predictor, was not a factor affecting the observed validity of the interview. Although demographic information was not available on the total sample, the breakdown on the 24 sales associates actually hired was 5 men, 19 women, 20 Whites, and 4 Blacks. The mean age was 29.7 years.

Criterion Measure and Final Form Internal Consistency

Approximately 9 months after the date of hire, sales productivity data for the 24 sales associates were taken from the company's records. In order to create a criterion measure, sales for each associate during each of the 20 pay periods reflected in the 9-month period were converted to sales-per-hour figures by dividing total sales volume by the number of hours worked during the same period. These 20 sales-per-hour ratios were then averaged to create an overall sales-productivity measure for each associate. Previous research conducted within the company on a much larger sample ($N = 573$) showed the odd-even reliability of this productivity measure to be .91 (computed by correlating the sum of the "odd" pay periods with the sum of the "even" pay periods over a 12-month period). The total interview score was created by summing the 16 item ratings (as computed across the 54 sets of ratings in this sample, Cronbach's alpha was .72).

Results and Discussion

Total interview scores for the 54 applicants ranged from 40 to 76 (out of a possible range of 16 to 80), with a mean of 57.3 and a standard deviation of 7.4. The validity coefficient between the total situational interview score and sales productivity was .45, which is significant at $p < .02$ (using a one-tailed test of significance; $N = 24$). In order to examine the predictive validity of the situational interview under ideal conditions, the observed coefficient was corrected for attenuation in the criterion ($r_{yy} = .91$). The validity of the situational interview, corrected for attenuation in the criterion, was .47.

In a recent meta-analysis of interview-validation studies, Hunter and Hunter (1984) found the mean validity of the employment interview to be .14. Obviously, the validity of the situational interview examined in the current research compares very favorably against this standard. The relatively high validity observed in the current study is somewhat surprising given the range restriction that may have existed in the total sample. Specifically, although applicants were apparently not selected on the basis of their interview scores, the initial sample was also probably not representative of the population of applicants for retail sales positions. All of these candidates had previous experience in retail jewelry sales, a fact that may explain the high mean interview score.

Even before the validation phase of this research was conducted, the situational interview was being implemented in stores across the country. (Because they previously worked for

the host department store, the interview was not known to the applicants in the validation phase of this study.) Follow-up conversations with managers using the form have revealed an overwhelmingly positive response to the situational interview. Most comments have focused on the job-relatedness of the interview, the ease of administration, and the ease of interpretation (many store managers had expressed a lack of confidence in their ability to assess a candidate's sales potential while avoiding potentially troublesome interview questions). However, like Latham and Saari (1984), we have found that the situational interview has not always been used as intended. For example, despite training efforts, a few managers have used the interview like an oral test, reading both the question and the anchors and then asking the applicant which of the anchors he or she thinks is best. Because it is different in appearance from other interview forms, researchers implementing a situational interview would be well advised to follow up with the user population after initial implementation to ensure that the procedure is being used correctly.

One limitation of this study, obviously, is the small sample size employed. Unfortunately, small samples seem to be the rule rather than the exception in interview research. Additionally, the interrater reliability of this interview is not truly known. Because the items were selected for inclusion on the basis of interrater reliabilities, the interrater reliability of the total form, as computed on the same data, is likely to be inflated. This probably explains why the interrater reliability of the form used here was slightly higher than that observed by Latham, Saari et al. (Latham & Saari, 1984; Latham et al., 1980). The internal consistency reliability of the situational interview was fair to good, depending on which set of data is considered. Internal consistency reliability computed on the pilot data ($r = .61$) was not as high as that estimated across the validation data ($r = .72$). It is possible that sample differences may account for the slight difference in internal consistency. At any rate, because items were selected partially on the basis of dissimilarity in terms of content (in order to sample as broadly as possible the domain of sales-associate performance), internal consistency reliabilities in the .60s were not surprising. One final limitation concerns the validity of the interview versus that of the interviewers. Because interview data were not coded for differences in interviewers, we were unable to determine if the situational-interview form was equally valid across both interviewers in the study or if, in fact, one of the interviewers was accounting for most of the observed validity.

Future research on the situational interview needs to expand into other higher-level jobs, particularly managerial positions. The usefulness of the procedure in these more complex jobs is not yet known. Further, it remains to be seen whether interviewers using a situational interview can accurately evaluate applicant potential on independent dimensions during an employment interview or whether global evaluations of potential are, as is usually the case in performance ratings, the best that can be done. Future research should be done in which a situational interview is constructed to reflect an a priori dimension structure (recall that because of implementation concerns, a simple unidimensional structure was adopted here). Until such research is done, the ability of interviewers to discriminate between performance potential on different job dimensions is un-

known. Currently, in any case, it appears that Latham, Saari, and colleagues have provided practitioners with a viable means of improving the overall reliability and validity of the employment interview.

References

- Arvey, R. D., & Campion, J. E. (1982). The employment interview: A summary and review of recent research. *Personnel Psychology, 35*, 281-322.
- Ghiselli, E. E. (1966). The validity of the personnel interview. *Personnel Psychology, 19*, 389-394.
- Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72-98.
- Janz, T. (1982). Initial comparisons of patterned behavior description interviews versus unstructured interviews. *Journal of Applied Psychology, 67*, 577-580.
- Latham, G. P., & Saari, L. M. (1984). Do people do what they say? Further studies of the situational interview. *Journal of Applied Psychology, 69*, 569-573.
- Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology, 65*, 422-427.
- Orpen, C. (1985). Patterned behavior description interviews versus unstructured interviews: A comparative validity study. *Journal of Applied Psychology, 70*, 774-776.
- Schmitt, N. (1976). Social and situational determinants of interview decisions: Implications for the employment interview. *Personnel Psychology, 29*, 79-101.
- Ulrich, L., & Trumbo, D. (1965). The selection interview since 1949. *Psychological Bulletin, 63*, 100-116.
- Wagner, R. (1949). The employment interview: A critical review. *Personnel Psychology, 2*, 17-46.

Received August 11, 1986

Revision received February 2, 1987

Accepted February 16, 1987 ■

Recognition of Facial Stimuli Following an Intervening Task Involving the Identi-kit

Sara Elizabeth Comish
University of Alberta, Edmonton, Alberta, Canada

The Identi-kit is a tool for constructing a facial composite. The types of errors made on a recognition task, following an intervening task involving the Identi-kit, were examined in this study. One hundred and eight introductory psychology students viewed a target composite-face and made an Identi-kit reconstruction. Subsequently, they were required to identify the original composite-face from a lineup of six composite-faces. Subjects who made an Identi-kit reconstruction were prone to make more errors on the recognition task than participants in a control condition, $p < .01$. False alarms were promoted when the subjects saw a lineup containing foils modified to resemble the subjects' own reconstruction errors, $p < .05$. When the foils in the lineup were not modified to resemble the subjects' own errors, those who made an Identi-kit reconstruction were more prone to miss the target, $p < .05$. This finding suggests that memory for facial stimuli can be influenced by viewing misleading information.

There is a commonly held belief that memory for faces is immune to interference; this belief is expressed by the oft heard statement, "I never forget a face." Support for the role of interference in facial memory, however, can be found in experimental studies. Deffenbacher, Carr, and Leu (1981) found that individuals' memory for photographs of faces was highly susceptible to retroactive interference, unlike their memory for nouns and objects, which was relatively immune to the interference. Loftus and Greene (1980) examined the extent to which facial memory was influenced by reading misleading descriptions of a previously shown target face. In a later recognition task with the target face absent, the majority of subjects who had read the misleading information picked a face with the misleading information. This finding raises the issue of whether facial memory can be altered by viewing misleading information.

A useful tool for studying this issue is the Identi-kit, a technique that consists of transparencies of line drawings of different facial features that can be superimposed on each other to make a composite-face. A similar system is the Photofit, which consists of actual photographs of features instead of line drawings. Studies that have used composite techniques to examine interference in memory for faces have yielded conflicting results. Mauldin and Laughery (1981) found that use of the Identi-

ti-kit to construct a target face improved recognition of the same target face in a later recognition task. On the other hand, Davies, Ellis, and Shepherd (1978) found different results. They found a trend in the opposite direction, which indicated that making a Photofit reconstruction could interfere with later recognition of a target face (although this difference failed to attain statistical significance). In both of these studies, subjects viewed a target face and then made a composite reconstruction. The recognition tasks that were used to assess memory involved photograph lineups that were not altered to resemble subjects' composite reconstructions.

It is reasonable to assume, however, that the source of any interference from Identi-kit reconstruction may result from viewing one's own errors in Identi-kit reconstructions as a holistic face and that maximum interference can be expected when the distractors used during the recognition task are similar to the individual's errors in Identi-kit reconstruction. The purpose of this study was to determine the effect of an intervening task involving misleading information on later recognition. In order to test this, subjects in two conditions made Identi-kit reconstructions of a target composite-face. The third condition was a control condition in which subjects did not make an Identi-kit reconstruction. For the subsequent recognition task, subjects saw a lineup consisting of the original target composite-face and five foils. The subjects were yoked on the recognition task to allow for the determination of whether interference was a result of making Identi-kit reconstructions per se or if it occurred only when the distractor foils resembled the participants' own Identi-kit reconstruction errors.

Method

Subjects

One hundred and eight introductory psychology students participated in partial fulfillment of a course requirement. Thirty-six subjects were randomly assigned to each of the three conditions.

Portions of this article were presented at the meeting of the Canadian Psychological Association held in Toronto, Canada, June 1986.

I am indebted to Gary Wells for the thoughtful help and guidance he provided during the course of this research. In addition, I am grateful to John Pullyblank, John Turtle, and Brendan Rule for their comments on an earlier version of this article; to Constable Bates of the Royal Canadian Mounted Police for his time and comments; and to two anonymous reviewers for their extremely helpful and meticulous comments.

Correspondence concerning this article should be addressed to Sara Elizabeth Comish, who is now at the Department of Psychology, University of Victoria, P.O. Box 1700, Victoria, British Columbia, Canada, V8W 2Y2.

Materials

The Identi-kit, available from Smith and Wesson, Inc., is a box containing transparencies of facial features. It is accompanied by the Identi-kit Handbook, which is a pictorial index of the features available in the kit.

Design and Procedure

A three-group design was employed. In the Identi-kit/own-errors condition, subjects made an Identi-kit reconstruction of a target composite-face and, in a later recognition task, encountered foils modified to resemble the subjects' own Identi-kit reconstruction errors. Subjects in the Identi-kit/other-errors conditions also made an Identi-kit reconstruction but were yoked to a subject in the Identi-kit/own-errors condition so that the foils in the recognition task were those of a subject in the Identi-kit/own-errors condition. In the control/other-errors condition, subjects saw the target composite-face but did not make an Identi-kit reconstruction. They were also yoked so that the foils in the recognition task were those of a subject in the Identi-kit/own-errors condition.

Initially, all subjects viewed one of two target composite-faces—photocopies of an Identi-kit reconstruction of either a man or a woman—for 10 s until it was removed. Subjects in the Identi-kit/own-errors condition and the Identi-kit/other-errors condition then made an Identi-kit reconstruction that involved an initial choice of facial features from the Identi-kit Handbook. A composite face was made using these features and shown to the subjects, who were allowed to change any of the features. On average, the reconstruction process took 13 min. Following this, subjects performed a filler task in which they were given 7 min to rate two photograph faces on trait dimensions. In order to ensure that the time that elapsed between the initial exposure to the target composite-face and the recognition task was the same for subjects in all conditions, the subjects in the control/other-errors condition spent 20 min rating the two photograph faces.

Finally, the subjects performed the recognition task, in which they were asked to select the original composite-face from a lineup consisting of the original target composite-face and five distractors. This task was not a forced choice and subjects were allowed to indicate that the target composite-face was not present in the lineup. In the Identi-kit/own-errors condition, each distractor foil was the same as the original target composite-face except that one error from the individual's Identi-kit reconstruction was substituted for one feature of each distractor. Thus, the first distractor was the same as the target except for the hair, the second distractor was the same as the target except for the nose, the third differed on the lips, the fourth differed on the eyes, and the fifth differed on the eyebrows (see Figure 1 for an example of a typical lineup). These distractor foils were photocopied and placed in a lineup, counterbalanced to vary the order of the target and the foils. Using this technique the target appeared in each position six times in each condition, with the positions of the foils randomly determined. Subjects in the Identi-kit/other-errors condition and control/other-errors condition saw the same lineup in the recognition task as the subject in the Identi-kit/own-errors condition to whom they were yoked.

The data from an additional four participants in the Identi-kit/own-errors condition and from an additional three subjects in the Identi-kit yoked condition were not included in the analyses because one or more features in the Identi-kit reconstruction was identical to the original target composite-face. When the correct feature was selected for the Identi-kit reconstruction, it was not possible to modify the lineup distractor foils in the manner described previously. These subjects were replaced.

The dependent variable was scored as a *hit*, a *miss*, or a *false alarm*. A hit was a correct recognition of the original target composite-face; a miss was an incorrect "not present" decision; and a false alarm was an

incorrect selection of a distractor foil. In addition, subjects' satisfaction with their reconstruction and the subjects' certainty with regard to their recognition were obtained on scales from *not at all satisfied* (1) to *totally satisfied* (7) and from *not at all certain* (1) to *totally certain* (7), respectively.

Results

All data were collapsed over the two target composite-faces and converted to percentages, which are presented by condition in Table 1. A one-way analysis of variance (ANOVA) across the three conditions was performed on the number of overall errors. A significant difference was found for the overall number of recognition errors, $F(2, 105) = 4.86, p < .01$. A Newman-Keuls comparison revealed that the control/other-errors condition produced significantly fewer errors than either of the Identi-kit reconstruction conditions. The estimated ω^2 was .065.

A further one-way ANOVA was performed on the number of false alarms. A significant difference was found among the conditions, $F(2, 105) = 4.46, p < .05$. A Newman-Keuls comparison revealed that the Identi-kit/own-errors condition differed significantly from the other two conditions, demonstrating that when the foils resembled subjects' Identi-kit errors, subjects were more prone to incorrectly identify the modified foils. In this case, the estimated ω^2 was .059.

Although an ANOVA on the number of overall errors and on both of the types of errors is redundant—because the two types of errors sum to give the number of overall errors—the effect of Identi-kit use on the different types of errors is of particular interest in this study. Therefore, an additional ANOVA was performed on the number of "not present" responses (misses). There was a significant difference among the conditions, $F(2, 105) = 3.75, p < .05$, and the Newman-Keuls comparison revealed that the subjects in the Identi-kit/other-errors condition made significantly more "not present" responses than subjects in either the Identi-kit/own-errors condition or the control/other-errors condition. The estimated ω^2 was .047.

An ANOVA was performed on subjects' confidence ratings, and there were no significant differences among the conditions, $F(2, 105) = 1.46, ns$. Wells and Lindsay (1985) have suggested that confidence ratings of "choosers only" (i.e., those subjects that make a recognition choice rather than those that indicate that the target composite-face was not present) should be analysed separately to determine whether confidence ratings have predictive use. No significant differences were found among the conditions when a one-way ANOVA was performed on the confidence ratings of "choosers only," $F(2, 89) = 0.422, ns$. The low accuracy–confidence correlation, however, is consistent with previous eyewitness studies (see Wells & Murray, 1984).

A point-biserial correlation between confidence and accuracy was performed for "choosers only" by condition and was not significant, highest $r = -.221$. Further, it was not significant when performed on the data of all subjects, highest $r = -.24$. A point-biserial correlation between subjects' satisfaction with their Identi-kit reconstruction and accuracy of recognition was also not significant for all subjects ratings nor was it significant when the analysis was performed on "choosers only," highest $r = -.27$.



Target Face



Reconstruction



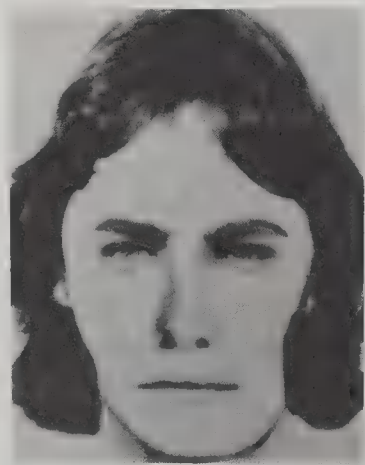
Foil - Hair



Foil - Nose



Foil - Lips



Foil - Eyes



Target Face



Foil - Eyebrows

Figure 1. The original target face, a typical Identi-kit reconstruction, and the subsequent lineup. (Foils are identical to the original target face, except for the indicated facial feature obtained from the reconstruction.)

Discussion

The use of the Identi-kit itself promotes overall errors on the recognition task. The lineup foils in this study were modified

by only one feature to resemble the Identi-kit reconstructions. This was sufficient, however, to bias people in favor of foils that resembled their own Identi-kit reconstructions. When the lineup foils resemble people's own Identi-kit reconstructions,

Table 1
*Performance on the Recognition Task: Data Converted
 Into Percentages by Condition*

Performance	Condition		
	Identi-kit (own errors)	Identi-kit (yoked)	Control (yoked)
Total errors	86.11 _a	77.78 _a	55.56 _b
Misses	8.33 _a	27.78 _b	8.33 _a
False alarms	77.78 _a	50.00 _b	47.22 _b
Correct (hits)	13.89	22.22	44.44

Note. Percentages that do not share a common subscript differ at $p < .05$.

the errors that are promoted are false alarms to composite-faces that resemble their Identi-kit reconstructions. Errors are promoted even when the composite-faces in the lineup are not similar to the subject's original Identi-kit reconstruction. In these cases, the predominant errors are misses, that is, failures to correctly identify the original target composite-face by saying that it is not present. Therefore, it appears that people who make an Identi-kit reconstruction are biased away from the original target composite-face.

The findings in this study appear to conflict with the results obtained by Davies et al. (1978) and Mauldin and Laughery (1981). The differences between the present study and the experiments of Davies et al. and Mauldin and Laughery are, however, numerous. Their null results, for example, may be due to a failure to allow the subject to make a "not present" response. Alternatively, the high similarity between distractors and target achieved in the present study, compared with the others, might account for the different results. Most important, no study prior to this has used Identi-kit reconstruction errors to build distractor composite-faces in which the target is embedded. This contingency between Identi-kit reconstructions and the recognition task in the Identi-kit/own-errors condition allowed for an optimal test of the potential interference effects of the Identi-kit task on subsequent recognition.

What is the source of the recognition errors? It is possible that the Identi-kit task creates interference because it involves viewing the misleading errors incorporated into a complete face. A person making a recognition decision may be easily confused between the original memory and the memory of the Identi-kit reconstruction. This is particularly likely to occur when a composite-face similar to a person's own Identi-kit reconstruction is present in the lineup.

The distribution of facial features in the foils that subjects had falsely identified was also examined. In general, people seem to make more recognition errors when features of less importance, such as eyebrows and lips, were altered. This is consistent with previous studies (see Ellis, 1984, for a review). Recognition errors involving the more important features, hair and

eyes, were primarily made only when the altered features resembled the Identi-kit reconstruction. This suggests that more confusions occur for less important features and for important features that resemble Identi-kit reconstructions.

Although the Identi-kit task was shown to interfere with recognition in this study, there are a number of caveats to consider in generalizing these findings to police lineups. First, it is not known how representative Identi-kit composite-faces are of photographs of real individuals, let alone the extent to which Identi-kit encoding operations compare with the encoding of live faces. Thus, performance on an Identi-kit lineup task may not be indicative of performance on a real lineup task. Second, the lineups used in this experiment consisted of highly similar composite-faces, faces much more similar than a typical police lineup. In the first condition, the foils were altered so that one feature was precisely the same as the subjects' Identi-kit reconstruction—an occurrence that is highly unlikely in a typical police lineup. Third, the target stimuli in this study were limited to only two Identi-kit composites that may not be representative of other faces. Fourth, the reconstructions from which the lineups were created shared no features with the target composite. Finally, the subjects in this study may have had poorer memories for the target than aroused witnesses to a crime and thus made poorer reconstructions. For these reasons, it is necessary to be cautious in extrapolating from these results to police lineups.

References

- Davies, G. M., Ellis, H. D., & Shepherd, J. W. (1978). Face identification: The influence of delay upon accuracy of Photofit construction. *Journal of Police Science and Administration*, 6, 35-42.
- Deffenbacher, K. A., Carr, T. H., & Leu, J. R. (1981). Memory for words, pictures, and faces: Retroactive interference, forgetting, and reminiscence. *Journal of Experimental Psychology: Human Learning and Memory*, 7, 299-305.
- Ellis, H. D. (1984). Practical aspects of face memory. In G. L. Wells & E. F. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives* (pp. 12-37). New York: Cambridge University Press.
- Loftus, E., & Greene, E. (1980). Warning: Even memory for faces may be contagious. *Law and Human Behavior*, 4, 323-334.
- Mauldin, M., & Laughery, K. (1981). Composite production effects on subsequent facial identification. *Journal of Applied Psychology*, 66, 351-357.
- Wells, G. L., & Lindsay, R. C. L. (1985). Methodological notes on the accuracy-confidence relation in eyewitness identifications. *Journal of Applied Psychology*, 70, 413-419.
- Wells, G. L., & Murray, D. M. (1984). Eyewitness confidence. In G. L. Wells & E. F. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives* (pp. 155-170). New York: Cambridge University Press.

Received July 20, 1986

Revision received November 19, 1986

Accepted October 20, 1986 ■

Instructions to Authors

Articles submitted for publication in the *Journal of Applied Psychology* are evaluated according to the following criteria: (a) significance of contribution, (b) technical adequacy, (c) appropriateness for the journal, and (d) clarity of presentation. In addition, articles must be clearly written in concise and unambiguous language. They must be logically organized, progressing from a statement of problem or purpose, through analysis of evidence, to conclusions and implications.

Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (3rd ed.). Articles not prepared according to the guidelines of the *Manual* will not be reviewed. All manuscripts must include an abstract of 100–150 words typed on a separate sheet of paper. Typing instructions (all copy must be double-spaced) and instructions on preparing tables, figures, references, metrics, and abstracts appear in the *Manual*. Also, all manuscripts are subject to editing for sexist language.

Authors can refer to recent issues of the journal for approximate length of regular articles. (Three double-spaced manuscript pages equal one printed page.) A few longer articles of special significance are occasionally published as monographs. Short Notes feature brief reports on studies such as those involving some methodological contribution or important replication.

APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more journals. APA policy also prohibits duplicate publication, that is, publication of a manuscript that has already been published in whole or in substantial part in another journal. Also, authors of manuscripts submitted to APA journals are expected to have available their raw data throughout the editorial review process and for at least 5 years after the date of publication.

Authors will be required to state in writing that they have complied with APA ethical standards in the treatment of their sample, human or animal, or to describe the details of treatment. (A copy of the APA Ethical Principles may be obtained from the APA Ethics Office, 1200 17th Street, N.W., Washington, DC 20036.)

Anonymous reviews are optional, and authors who wish anonymous reviews must specifically request them when submitting their manuscripts. Each copy of a manuscript to be anonymously reviewed should include a separate title page with authors' names and affiliations, and these should not appear anywhere else on the manuscript. Footnotes that identify the authors should be typed on a separate page. Authors should make every effort to see that the manuscript itself contains no clues to their identities.

Manuscripts should be submitted in quadruplicate and all the copies should be clear, readable, and on paper of good quality. A dot matrix or unusual typeface is acceptable only if it is clear and legible. Authors should keep a copy of the manuscript to guard against loss. Mail manuscripts to the Editor, Robert Guion, Department of Psychology, Bowling Green State University, Bowling Green, Ohio 43403.

Journal of Applied Psychology Monograph

Meta-Analysis of Assessment Center Validity

Barbara B. Gaugler, Douglas B. Rosenthal, George C. Thornton III, and Cynthia Bentson
Colorado State University

Meta-analysis (Hunter, Schmidt, & Jackson, 1982) of 50 assessment center studies containing 107 validity coefficients revealed a corrected mean and variance of .37 and .017, respectively. Validities were sorted into five categories of criteria and four categories of assessment purpose. Higher validities were found in studies in which potential ratings were the criterion, and lower validities were found in promotion studies. Sufficient variance remained after correcting for artifacts to justify searching for moderators. Validities were higher when the percentage of female assesseees was high, when several evaluation devices were used, when assessors were psychologists rather than managers, when peer evaluation was used, and when the study was methodologically sound. Age of assesseees, whether feedback was given, days of assessor training, days of observation, percentages of minority assesseees, and criterion contamination did not moderate assessment center validities. The findings suggest that assessment centers show both validity generalization and situational specificity.

Since the first industrial application of assessment centers in 1956 by AT&T (Bray & Grant, 1966), a growing number of companies have used the assessment center method. Today, it is estimated that more than 2,000 organizations are currently using some type of assessment center program. Organizations use assessment centers for a wide variety of purposes, including selection, placement, early identification of management potential, promotion, development, career management and training. Although assessment centers are most frequently used for assessing managers, they have also been developed to assess college students, engineers, salespersons, military personnel, rehabilitation counselors, school administrators, and blue-collar workers.

The increasing popularity of the assessment center method has stimulated a great amount of research concerning its effectiveness. Reviewers have accumulated research findings from a variety of types of assessment centers and have concluded that assessment centers have predictive validity for a variety of criteria (Byham, 1970; Cohen, Moses, & Byham, 1977; Howard, 1974; Thornton & Byham, 1982). Although the predictive validity coefficients of assessment centers are generally high, some assessment centers have low predictive validity. In fact, the observed validity coefficients of the assessment centers reviewed by us ranged from $-.25$ to $+.78$. The current meta-analysis was

designed to estimate the true validity of assessment centers and to understand the reasons for the variability in observed predictive validity coefficients.

Meta-analysis is a collection of methods used to aggregate results across studies quantitatively. It helps us draw more accurate conclusions about inconsistent findings in a particular area of research. Statistical procedures replace the traditional literature review, which has been criticized for its "great information-gate-keeping potential" (Cooper & Rosenthal, 1980, p. 442). Literature reviews are highly influenced by the biases of the reviewer, may neglect large amounts of information provided in the original research reports, and imprecisely weight conclusions with regard to the amount of research covered. Statistical techniques of aggregation have been suggested as an alternative to the literature review.

There are a number of different meta-analytic procedures that vary in sophistication (Rosenthal, 1978). The method developed by Schmidt and Hunter (1977) was used in this study for a variety of reasons. First, it provides specific formulas for statistically cumulating effect sizes across studies. Second, it was developed specifically for use with correlational data (e.g., validity coefficients). Third, it rests on the assumption that much of the variation in observed results is due to statistical artifacts and methodological problems rather than to true differences in underlying population correlations. Artifacts include sampling error due to studies having sample sizes less than infinity, unreliability of predictor and criterion measurement, differential restriction of range across samples, and various typographical and other data analysis and reporting errors. Formulas are used to estimate the amount of true variation in validity coefficients and the amount of observed variation that is due to artifacts.

Validity generalization studies of other selection techniques suggest that a substantial amount of the variability in predictive validities is due to statistical artifacts. Studies of programmer and clerical aptitude tests (Pearlman, Schmidt, & Hunter, 1980;

Parts of this article were presented at the 93rd Annual Convention of the American Psychological Association held in August 1985 in Los Angeles.

We would like to acknowledge the helpful comments of three anonymous reviewers.

Douglas B. Rosenthal is now at HUMRRO in Alexandria, Virginia. Cynthia Bentson is now at Bentson Associates, Fort Collins, Colorado.

Correspondence concerning this article should be addressed to Barbara B. Gaugler, who is now at Rice University, Psychology Department, Houston, Texas 77251.

Schmidt, Gast-Rosenberg, & Hunter, 1980), mechanical and chemical comprehension tests (Schmidt, Hunter, & Caplan, 1981), weighted biographical inventories (Brown, 1981), tests of general intelligence (Pearlman et al., 1980; Schmidt et al., 1981), and also verbal, quantitative, reasoning, spatial/mechanical and motor ability, perceptual speed, memory, and performance tests (Pearlman et al., 1980) have shown that at least 60% of the variation in single predictor-criterion relations can be accounted for by sampling error, predictor and criterion unreliability, and range restriction.

The predictive validity of assessment centers may be susceptible to the same artifacts. First, sampling error contributes to variability in validity coefficients. Because of the expense and time required by the assessment center process, many of the studies of the criterion validity of assessment centers have relatively small sample sizes. Thornton and Byham (1982) reported that sample sizes varied from 12 to 5,943, with a median of 55. Most studies used 40 to 50 candidates, and only a few studies had over 100 subjects. Because sampling error accounted for most of the artifactual variance in the previous validity generalization studies, and because sample sizes of assessment center studies are relatively small and vary to such an extent, it was predicted that sampling error would account for much of the variance in assessment centers' predictive validity coefficients.

Second, assessment center studies show moderate to severe levels and variation in range restriction. When assessment center results are used for operational purposes, not everyone who is assessed is selected or promoted.

Reliability of supervisory ratings of performance and potential, a common criterion in assessment center studies, may be low. As Thornton and Byham (1982) noted, "Problems with supervisors' ratings are legion. Leniency, halo, and restriction-in-range biases may occur" (p. 298). In addition, the reliability of the criterion measured to validate assessment centers varies considerably. For example, the mean reliability of the criteria reviewed for this meta-analysis ranged from .61 to 1.00.

Although there is reason to believe that variability in validity coefficients of assessment centers may be partially due to methodological artifacts, the diversity in makeup of assessment centers suggests the possibility that certain variables may moderate the predictive validity of assessment centers. There is such a variety of assessment center procedures that a *typical* assessment center does not exist (L. Alexander, 1979; Bender, 1973; Byham, 1978a, 1978b; Thornton & Byham, 1982). "There is no standardization of content in assessment centers or of the way they are administered, and there is no uniform method of treating the performance evaluation data generated by assessment centers" (Bender, 1973, p. 56). Such variety makes informed comparisons across studies extremely difficult. Meta-analysis provides a quantitative method of examining correlates of the predictive validity coefficients.

Many of the individual investigations of moderators of assessment centers, using single samples, have yielded conflicting and inconclusive evidence. Moderators that have been tested in individual studies include the following: candidate's age (Burrroughs, Rollins, & Hopkins, 1973; Neidig, Martin, & Yates, 1978), candidate's minority group status (S. Alexander, 1975; Clingenpeel, 1979; Huck, 1974; Huck & Bray, 1976; Jaffee, Cohen, & Cherry, 1972; Marquardt, 1976; Moses, 1973a, 1973b; Moses & Boehm, 1975; Russell, 1975), sex of candidate (S. Al-

exander, 1975; Clingenpeel, 1979; Hall, 1976; Marquardt, 1976; Moses, 1973a, 1973b; Moses & Boehm, 1975), composition of the assessee group (Byham, 1981; Schmitt & Hill, 1977), type of criterion (Klimoski & Strickland, 1977, 1981), and time at which criterion measures are taken (Finley, 1970; Hinrichs, 1978; A. Howard, personal communication, February 16, 1979; Mitchel, 1975; Moses, 1972; Slivinski & Bourgeois, 1977). In addition, other parameters have been suggested for investigation (Thornton & Byham, 1982): types of evaluation devices, operating procedures, ratio of assessees and assessors, evaluation of observed dimensions, process of integrating information, uses made of performance data, and purposes of assessment. Meta-analysis provides a method of examining variability in validity coefficients *across* studies of different populations.

In the present investigation, both validity generalization (i.e., whether the lower bound of some confidence interval around the average validity is greater than zero) and situational specificity (i.e., whether nonartifactual variance in validities exist) were studied. It is possible and meaningful to find any combination of results. Hunter, Schmidt, and Jackson's (1982) meta-analytic procedures were applied to the results of 50 studies that investigated the relation between the overall assessment rating and various criteria. The purpose was threefold: (a) to estimate the true validity of assessment centers, (b) to determine the extent to which varied results across studies are due to statistical artifacts and methodological problems, and (c) to discover which characteristics of assessment centers moderate the predictive validity of assessment centers.

Method

Selection of Studies

A review of the literature was undertaken using *Psychological Abstracts*, reference lists of previous reviews (Cohen et al., 1977; Howard, 1974; Huck, 1973; Thornton & Byham, 1982), and personal contact with primary researchers in the field. From this pool, published and unpublished studies were selected to be included in the meta-analysis that met the following criteria: (a) The manuscript described an assessment center, as delineated by the Standards for Assessment Centers (Task Force on Assessment Center Standards, 1980), and (b) a correlation between the overall assessment rating and some criterion was provided or calculable from the data given. Studies included experimental studies in which there was no operational use of assessment center data, studies with no feedback to participants, studies that compared the subsequent performance of assessed and nonassessed groups (i.e., control groups), correlational studies with feedback to assessees and management, and concurrent validity studies. No study was excluded on the basis of poor method or quality. However, the quality of various design features, adequacy of information provided, and external validity of each study were rated by the authors.

Ratings of Characteristics

A number of variables were believed on theoretical or empirical grounds, or both, to contribute to the relation between the overall assessment rating and various criteria. We also examined demographic and other variables of interest. The following information was recorded for each study: (a) identification information—coder identification number, study identification number, effect size identification number, publication year, publication form, and country of study; (b) candidate characteristics—average age at time of assessment, educational level,

current position, percentage of men, and percentage with minority status; (c) assessment center description—types of assessment techniques used (e.g., in-basket, leaderless group discussion), number of types of assessment techniques, ratio of assesseees to assessors, names of dimensions assessed, process of integrating information, uses made of assessment data, purposes of assessment, types of criteria, and time lag between the assessment center and when the criterion was measured; (d) study design and reporting—study design (e.g., experimental), number of assesseees, number of assesseees on whom criterion measures were taken, reliability of the overall assessment rating, reliability of the criterion, presence of a systematic method of identifying dimensions, presence of potential restriction of range, potential for criterion contamination through knowledge of assessment results, threats to the validity of research (e.g., inadequate representation; Cook & Campbell, 1976), general index of validity of research, and adequacy of information provided in the report; and (e) conclusions—uncorrected correlation value, statistic given, and author's conclusion about calculated correlation.

The four authors of this article served as coders of the characteristics of the studies. Following training and practice, interjudge agreement in coding among the four authors was calculated, once prior to and once during the coding of the studies. Because the characteristics were coded on a variety of scales of measurement, several indices of interrater reliability were needed. Acceptable levels of interrater agreement (i.e., >85%) were attained for all types of variables for both assessments. For example, for categorical variables, there was total agreement on 78% of the items, agreement among three judges on 11% of the items, and agreement by two judges on 11% of the items. Kuder-Richardson 20 reliabilities performed on the dichotomous items averaged .91 and .88 for the two interrater reliability assessments. Detailed data are available from the first author.

Analytic Procedures

Combining validities within studies. Many of the studies reported multiple validity coefficients. In some cases, researchers obtained several independent samples of assesseees and calculated separate validities on each sample. Validities from multiple samples were considered statistically independent and were therefore entered unchanged into the cumulation formulas. (Validities for multiple samples may not be independent because of similarities in exercises, biases of assessors, biases of supervisors providing criterion ratings, and generalized features of the organization climate. Supplementary analyses were conducted that combined validities judged to be dependent within a research report.) More frequently, however, researchers used several criterion measures for the same sample (e.g., supervisor performance ratings, salary advancement, number of promotions). Validities calculated on the same sample were considered statistically dependent and therefore were combined, following the recommendations of Hunter et al. (1982, p. 118). In most cases, a simple mean was calculated across dependent validities within a single study. In a few cases, intercorrelations among criteria were reported that allowed us to compute a composite validity (Hunter et al., 1982, p. 120). The advantage of using a composite validity instead of a mean validity is that it reflects the validity the researchers would have obtained, had they originally summed each assessee's criterion scores, and then correlated them with overall assessment ratings. Dependent validities were only combined within five general categories of criteria. Table 1 presents the five categories and the individual criteria subsumed under each.

Some studies reported multiple validities on the same sample from a single criterion measured at various times (e.g., number of promotions in the 1 year, 5 years, and 10 years since the assessment center). In a preliminary analysis, a small, nonsignificant correlation between the time when criteria were obtained and the magnitude of validities was found. (Concurrent studies were excluded from this analysis.) We therefore decided to combine all validities taken at different times for the

Table 1
Five Categories of Criteria

Category	Criteria within category
Rating of job performance	An overall performance rating Field observation of manager's performance Field interview with supervisor of the manager A rating on some aspect of job performance other than an assessment center dimension
Potential ratings	A rating of manager's potential
Dimension ratings	Rating of manager's job performance on the dimensions used in the assessment center
Performance in training	Performance of manager in a training program
Career advancement	Change in salary over time Absolute level of salary obtained Number of promotions Absolute job level obtained Turnover

same criterion category within studies. Because subject attrition over time generally occurred in these studies, a mean sample size was calculated for each study.

Cumulating effect sizes across studies. First, the mean validity and variance of validities weighted by sample size were calculated. Thus, large studies are given more importance than small ones. Second, the mean and variance were corrected for statistical artifacts. Using distributional formulas presented by Hunter et al. (1982, p. 90), the weighted mean validity was corrected for restriction of range and unreliability in the criteria.¹ The validities were not corrected for predictor unreliability because we were unable to obtain a reasonable estimate of the distribution of reliabilities for the overall assessment ratings, and to correct for unreliability in the overall assessment rating would yield a mean validity that assumes overall assessment ratings are perfectly reliable. Such a mean would be an overestimate of the validity of assessment centers as currently practiced. It is important to note that variability in the unreliability in the overall assessment ratings is an artifactual source of variance in assessment center validities and, ideally, should be removed from the variance. However, we could not do this without reliability estimates of the overall assessment ratings. The correction formulas used in this study are the same correction formulas used by Schmidt, Hunter, and their colleagues in some of their early validity generalization work on personnel selection research (see Hunter et al., 1982, p. 91). In these validity generalization studies the selection tests were treated as fixed (i.e., variance due to predictor unreliability was ignored).

Table 2 presents the criterion reliability estimates used to correct the validity means and variances. Means and variances of the square roots of the reliabilities are presented because they are the actual numbers used in the correction formulas. It was assumed that the reliability distributions for performance ratings, ratings of potential, and dimension ratings would be identical because they are all ratings of on-the-job performance. This reliability estimate was calculated by combining the reliabilities for performance, potential, and dimension ratings reported in our assessment center studies and reliabilities from other research on performance evaluation. The result was a list of 286 reliabilities for which means and variances were computed.² Our estimate of reliability

¹ A typographical error appears in Hunter, Schmidt, and Jackson (1982, p. 90) for the formula to compute "*c*". It should read \bar{r}^2 , not \bar{r} . The analyses appearing in the present article were computed using the correct formula.

² These are available from the second author.

Table 2
Means and Variances of Assumed Reliability Distributions

Criterion	Mr_{yy}	$M\sqrt{r_{yy}}$	$\sigma_{r_{yy}}^2$	$\sigma^2\sqrt{r_{yy}}$
Performance, potential, and dimension ratings	.61	.774	.034	.016
Performance in training	.80	.894	.007	.002
Career advancement	1.00	1.000	.000	.000
Total sample	.77	.871	.039	.015

Note. Mr_{yy} = estimate of mean criterion reliability; $M\sqrt{r_{yy}}$ = estimate of the mean of the square roots of criterion reliabilities; $\sigma_{r_{yy}}^2$ = variance of criterion reliabilities; $\sigma^2\sqrt{r_{yy}}$ = variance of the square roots of criterion reliabilities.

ties for training criteria came from an assumed distribution reported by Pearlman et al. (1980, p. 375). For the criterion measures in the career advancement category, we assumed a mean validity of 1 and a variance of 0 because we could find no relevant data and we wished to use a conservative (i.e., high) figure. To obtain the reliability values for the total sample of studies, a distribution was created using the reliability estimates from each category in proportion to the number of studies we had in each category. Means and variances were then calculated from this distribution.

Moderator analyses. Moderator analyses were not undertaken until it was determined that enough variance in correlations remained after correcting for statistical artifacts to warrant such a search (Hunter et al., 1982). In all, 20 potential moderators were tested. Continuous and dichotomous moderators were tested by correlating them with study validities. Although other studies have tested dichotomous moderators by dividing validities into groups and comparing corrected means and variances, we chose to compute point-biserial correlations because this gave us statistics comparable to the other moderator analyses using correlations. We did however, test potential moderators with three or more categories by comparing corrected means and variances of each group of validities.

Several factors influenced our choice of variables to test as potential moderators. First, there had to be sufficient variance on the variable to allow a meaningful test. Second, the variable had to meet one of the following criteria: Past research suggested its moderating effects, the results would have relevance to concerns for equal employment opportunity,

Table 3
Means and Standard Deviations for Variables Tested as Moderators of Assessment Center Validities

Variable	<i>N</i>	<i>M</i>	<i>SD</i>
Publication year	108	74.00	7.77
Mean age of assesseees	57	30.15	6.89
Percentage men	68	63.98	43.42
Percentage minority status	37	15.69	28.83
Total no. of devices	104	7.33	1.99
Days of assessor training	67	7.52	5.27
No. of hours per assessee spent integrating information	53	1.62	.52
Quality of the study as measured by an overall judgment made by the authors of this article	108	2.14	1.00
Quality of the study as measured by summing ratings of threats to validity	105	1.68	1.52

Table 4
Descriptive Information for Variables Tested as Moderators of Assessment Center Validities

Variable	<i>N</i>	Variable	<i>N</i>
Publication form	109	Days of observation of assesseees	96
Published	63	1	17
Journal	60	2	57
Book	2	3 or more	22
Thesis	1	Psychologists vs. managers as assessors	76
Unpublished	21	Psychologists	10
Used an intelligence test	106	Managers	66
Yes	78	Use of peer evaluations	93
No	28	Yes	44
Ratio of assesseees to assessors	80	No	49
1:1	6	Feedback given to assesseees	87
2:1	57	Yes	50
3:1	11	No	37
4 or more:1	6	Feedback given to immediate supervisor	77
		Yes	16
		No	61

or the results might have practical relevance to the design or administration of assessment centers.

Tables 3, 4, 5, 7, and 11 list the potential moderators. Additional comment is warranted on the last two variables, type of criterion and purpose of the assessment. It was felt that differences in criterion type and purpose of the assessment center implied conceptually distinct types of validity information. It was also suspected that the other 18 potential moderators might operate differently within the categories of these two variables. So, at one point in our analyses, validities were sorted on the basis of criterion type and assessment purpose and then were tested for moderators within each of these sortings.

Large-sample studies. The distribution of sample size in our studies was positively skewed due to the presence of three relatively large studies. Moses and his colleagues (Moses, 1972; Moses & Boehm, 1975; Ritchie & Moses, 1983) used samples of 5,943, 4,846, and 1,097 assesseees, respectively. These samples are substantially larger than the next largest sample, which contained 471 assesseees. Because there is a chance that weighted means and variances could be misleading when samples of this magnitude exist (Hunter et al., 1982, p. 41), our meta-analysis was carried out twice, once including the disparate studies and once excluding them. Within the total sample of studies, only a small, nonsignificant decrease was found in the corrected mean and variance when the large studies were removed. However, the three large studies were excluded from subsequent calculations in our meta-analysis because it was suspected that they would predominate in the subgroup analyses that contained fewer studies.

Results

Descriptive Information

Table 6 lists a number of the characteristics of the studies in our meta-analysis. There is wide diversity in the design of assessment centers and their predictive validity studies. Among the studies (28% of the total) that reported minority status, on the average 17% of the assesseees were minorities. The total number of different types of assessment devices ranged from 1 to 11 with a mean of 7 per study. The number of days of obser-

Table 5
Ratings of Quality of Studies Tested as Moderators of Assessment Center Validities

Ratings on individual threats	No plausible threat	Minor problems	Plausible threat	Could explain most of results
Inadequate representation ^a	41	30	37	
Motivation differences ^a	96	4	7	
Job experience ^a	78	21	8	
Criterion contamination	50	25	25	5
Other	91	4	9	1

^a Summed to form a composite rating of study quality.

vation ranged from one to three days. For most studies (64%), managers served as assessors; some employed both psychologists and managers (20%), whereas a few used only psychologists as assessors (10%).

Most studies were published in journals (52%), others were presented at conferences (22%) or were prepared as in-house technical reports (22%). The plurality of the assessment centers reviewed were conducted for promotional purposes (46%); however, others were carried out for the purpose of selection (22%), developmental planning (4%), early identification of managerial talent (16%), or basic research (6%). Most of the studies (52%) used a predictive validation design and provided feedback to assesseees regarding their performance in the assessment center. Others used a predictive design but did not provide feedback (19%), were pure research experiments (16%), used a control group (4%), or used a concurrent validation design (20%). In addition, many studies either used job performance ratings ($n = 28$) or measures of career advancement ($n = 25$) as criteria, whereas others used ratings of potential ($n = 9$), measures of performance in training ($n = 7$), or dimension ratings ($n = 5$). The variety of assessment centers described here substantiates the contention that a typical assessment center does not exist.

The 50 studies reported 220 validities that when combined within criterion types within each study resulted in 112 validity coefficients. Three large-sample studies and two studies that failed to report sample size were excluded from our analyses, yielding a final total of 107 validity coefficients. Fifteen studies contributed approximately one half of these validities, but several were computed from independent samples of subjects. Supplementary analyses were conducted after additional combinations of potentially dependent validities.

Test for Validity Generalization

Cumulation of effect sizes across studies. Table 7 presents unweighted means and variances for the total set of validities and for validities sorted into categories of criteria type and purpose of the assessment center. The unweighted mean validity (\bar{r}) across the total sample was .32. Most of the mean validities within the criteria and purpose categories were close to this value. The two exceptions were the validities for studies using ratings of potential as the criterion ($\bar{r} = .45$), and for research studies ($\bar{r} = .42$). None of the mean validities changed significantly when they were recalculated, including the three outliers.

The last column of this table contains the raw or unweighted variances (s_r^2) across validities. The s_r^2 s of most categories

ranged from .03 to .04. However, validities using dimension ratings as the criterion had a relatively large variance (.071), whereas assessment centers conducted for the purposes of research, early identification, and career advancement had relatively little variance.

Correction for artifacts. Table 8 presents the weighted mean validities and variances corrected for statistical artifacts. According to Hunter et al. (1982), weighting by sample size increases the accuracy of population estimates. The relative magnitude of the weighted and unweighted means and variances are quite similar. However, in all cases, the weighted means and variances (Table 8) are slightly lower than the unweighted values (Table 7). For example, the unweighted mean and variance across the total set of validities are .32 and .030, respectively. Weighting by sample size reduces these numbers to .29 and .023. These reductions result from a negative correlation, $r = -.24$ ($p < .05$), between sample size and size of validities.

The column headed by $\bar{\rho}_{xU}$ presents the weighted means corrected for range restriction and unreliability in the criterion.³ The corrected mean across the total sample of validities is .37. The corrected means within the criterion categories are about .35, with the exception of the average validity for studies using ratings of potential, which had a noticeably larger mean of .53. In the purpose categories, early identification, selection, and research studies have corrected mean validities between .41 and .48, whereas mean validity for promotion studies was somewhat smaller (.30).

These corrected means were computed by dividing the uncorrected weighted means by the product of the estimates of the average of the square root of the criterion reliabilities, $E(r_{yy}^{1/2})$, and values incorporating estimates of range restriction, \bar{c} . (See Hunter et al., 1982, p. 90.) Small c values indicate more range restriction than large c values.

A comparison of Column 1 with Column 3 reveals that the correction for unreliability and range restriction boosted some validities more than others. Within the criterion categories, performance, potential, and dimension means increased by at least .11. However, training and career progress mean validities increased by about only one half this amount. In studies in which training performance was the criterion, the smaller increase was due to the lack of range restriction (i.e., $\bar{c} = .977$). In studies

³ Using Hunter, Schmidt, and Jackson's (1982) notation, "U" represents both range restriction (\bar{U}) and corrected criterion scores in the notation of the mean validity corrected for artifacts ($\bar{\rho}_{xU}$).

Table 6 (continued)

Author(s)	Form of study ^a	% of minority	No. of devices	No. of days of assessor observation	Type assessor	Peer evaluation	Purpose of AC	Design of study	Criterion type	Validities used in the meta-analysis ^b
Hinrichs (1978)	Journal	NR	8	2	Managers	No	Early ID	Predictive (w/o feedback)	Career	.46
Howard (1979)	Presentation	0	7	3	Psychologists	Yes	Research	Experimental	Career	.33, .44
Huck & Bray (1976)	Journal	0, 100, 28	8	2	Managers	Yes	NR	Predictive (w/feedback)	Job performance	.41, .35, .40, .44
								Predictive (w/o feedback)	Potential	.59, .54, .52, .64
Klimoski & Strickland (1981)	Unpublished	NR	7	2	Managers	Yes	Promotion	Predictive (w/feedback)	Dimension	.47 ^c (6), .41 ^c (6)
Kraut & Scott (1972)	Journal	NR	10	2	Both	NR	Early ID	Predictive (w/feedback)	Job performance	.02
McConnell & Parker (1972)	Journal	33, 0	11	1	Managers	Yes	Promotion	Concurrent	Potential	.37
McElroy (1979)	Unpublished	6	NR	NR	Both	No	Promotion	Predictive (w/feedback)	Career	.22 (3)
									Career	.28, .29, .33, .30, .26, .27
									Job performance	.55, .48, .64, .28
									Job performance	.51, .41
Metropolitan Transit Authority (1972)	Unpublished	20	5	1	Managers	NR	Development Planning	Predictive (w/feedback)	Career	.52
Mitchel (1975)	Journal	NR	9	3	Managers	Yes	Promotion	Concurrent	Job performance	.71
Moses (1972) ^f	Presentation	NR	6	2	Managers	NR	Promotion	Predictive (w/feedback)	Career	.16 (2), .19, .25 (3)
Moses & Boehm (1975) ^f	Journal	NR	6	2	Managers	Yes	Promotion	Predictive (w/feedback)	Career	.44
Moses & Wall (1975)	Journal	77	4	1	Managers	NR	Promotion	Predictive (w/feedback)	Career	.25 (3)
Norton (1980)	Presentation	NR	4	NR	Managers	No	Selection	Predictive (w/o feedback)	Job performance	.60
Parker (1980)	Journal	6	11	1	Managers	Yes	Early ID	Concurrent	Job performance	.30
Ritchie (1980)	Presentation	20	4	1	Managers	NR	Promotion	Concurrent	Job performance	.36
								Experimental	Job performance	.10
									Training	.14
									Career	.03
Ritchie & Moses ^f (1983)	Journal	NR	5	2	Managers	No	Development Planning	Predictive (w/feedback)	Career	.42
Schmitt, Noe, Meritt & Fitzgerald (1984)	Presentation	15	6	2	Managers	No	Promotion	Predictive (w/feedback)	Job performance	.25, .29, .09
Slivinski, Grant, Bourgeois, & Pederson (1977)	Unpublished	NR	10	2	Managers	No	Promotion	Predictive (w/feedback)	Dimension	-.03 (7), -.04 (7)
									Job performance	.38 (3), .14 (3)
									Career	.42 (6), .29 (4)

Table 6 (continued)

Author(s)	Form of study ^a	% of minority	No. of devices	No. of days of assessor observation	Type assessor	Peer evaluation	Purpose of AC	Design of study	Criterion type	Validities used in the meta-analysis ^b
Slivinski, McCloskey, & Bourgeois (1979)	Presentation	NR	5	NR	NR	No	NR	Predictive (w/o feedback)	Career	.30
Thomson (1969)	Journal	NR	9	3	Managers Psychologists	Yes	Promotion	Predictive (w/feedback)	Potential Dimension	.64, .64 .57 ^c (12), .61 ^c (12)
Tziner (1982)	Unpublished Journal	NR	5	2	Managers	Yes	Selection	Experimental	Training	.35
Tziner (1984)	Journal	NR	4	2	Managers	Yes	Selection	Predictive (w/feedback)	Job performance Training	.30 .60
Tziner & Dolan (1982)	Journal	NR	6	2	Managers	No	Selection	Experimental	Training	.38
Vernon (1950)	Journal	NR	8	2	Both	No	Selection	Predictive (w/feedback)	Job performance ^b	.25, .13, .22, .16
Warriner (1981)	Presentation	NR	10	2	Managers	NR	Early ID	Predictive (w/feedback)	Training Career	.42 .39
Wilson (1948)	Journal	NR	8	2	Both	No	Selection	Predictive (w/feedback)	Job performance	.50
Wissman & Rankin (1982)	Presentation	NR	5	NR	Managers	No	Promotion	Predictive (w/feedback)	Job performance Dimension	<u>-.02</u> , <u>.0</u> .06 (14)
Wollowick & McNamara (1969)	Journal	NR	10	2	Managers	No	Early ID	Predictive (w/o feedback)	Career	.37
Worbois (1975)	Journal	8	11	1	Managers	No	Promotion	Concurrent	Job performance ^b	.38 (5)

Note. NR = not reported by the study. AC = assessment center. Double entries for % minority, type assessor, peer evaluation, purpose, and design reflect an article reporting multiple independent investigations. Double and triple entries for criteria type and validities used, reflects either multiple investigations or our desire to categorize criterion type into one of five types.

^a When a study appeared in a journal and in other forms (e.g., presentation), we recorded it only as a journal.

^b The numbers in parentheses indicate number of validities combined in the average or composite. Values that are underlined were combined in the supplementary analyses of potentially dependent validities.

^c Summed or averaged dimension ratings used as the criterion were recorded as a measure of job performance.

^d Rather than testing an average, we combined the validities into a composite.

^e J. J. Erpenbach (personal communication, March 11, 1971).

^f These studies were excluded from the bulk of the analyses because of their relatively large sample sizes.

Table 7
Unweighted Means and Variances of Validities

Sample	No. of studies	No. of validities	Sample range	Total sample	Unweighted \bar{r}^a	Unweighted s^2_r
Total	47	107	12–471	12,235	.32	.030
Criteria						
Performance	29	44	12–471	4,180	.31	.032
Potential	9	13	20–425	1,338	.45	.037
Dimension	5	9	35–122	748	.25	.071
Training	8	8	50–269	1,062	.31	.031
Career	22	33	30–437	4,907	.32	.011
Purpose ^b						
Promotion	21	52	13–53	5,201	.29	.040
Early ID	8	15	24–437	2,068	.31	.009
Selection	12	24	55–301	3,198	.30	.019
Research	3	6	125–144	837	.42	.003

Note. Unweighted \bar{r} = simple mean validity (i.e., not weighted by sample size). Unweighted s^2_r = simple variance of validities (i.e., not weighted by sample size).

^a The mean correlation coefficients did not change when we recalculated them including the three large-sample studies.

^b The total of the purpose categories fails to sum to 107 because we excluded two validities in a developmental planning category and were unable to classify several others.

with career progress measures, the smaller increase was due to the high reliability estimates for this criterion.

Similar observations can be made within the purpose categories. Mean validities of early identification and selection studies increased by at least twice the amount for promotion and research studies. The relatively smaller increases for promotion and research were due to lack of range restriction within these studies.

In column 4, $\sigma_{\rho_{xU}}^2$ represents the variances of the weighted validities corrected for all statistical artifacts. These values were computed using a formula presented by Hunter et al. (1982, p. 90). Schmitt, Gooding, Noe, and Kirsch (1984) noted that this correction may be inaccurate when applied to small samples. Thus, caution should be exercised when interpreting the corrected variances for the research and training studies, and studies using dimension ratings as criteria. In most categories, the correction reduced the original weighted variances. For career, early identification, selection, and research validities, most or all of the variance appears to have been artifactual. Partial support for this conclusion was found using a chi-square test developed by Hunter et al. (1982, p. 47), which was applied to uncorrected weighted variances. However, the results of the chi-square analyses should be interpreted with caution. Whereas nonsignificant results, as found for early identification, career progress, selection, and research validities, suggest that no true variance exists among these populations of validities, significant results are ambiguous and can be caused by artifacts, true variance, or both. In addition, even when significant results are based on true variances, the amounts may be trivial in size.

The last three columns provide information about the distributions of corrected validities. The lower 90% credibility value is the point above which lie 90% of the true validities. This statistic can be used to assess the likelihood that any given assessment center will be at least minimally valid. This value exceeds zero for all of the studies except those in which dimension ratings were used as criteria. The final two columns depict the lower and upper bounds for the 95% confidence interval created

around the corrected mean. Using this more stringent criterion, all categories of studies except those using dimension ratings and those conducted for promotional purposes, appear to be at least minimally valid.

Supplementary analyses were conducted to examine whether potentially nonindependent validities within studies affected the results of this meta-analysis. Each study in Table 6 that contained multiple validities was reexamined. A judgment was made by the third author about whether the separate coefficients were potentially dependent. Validities were considered potentially dependent and, thus, were combined (a) if it appeared that the same or quite similar assessors were involved, (b) if the study was done in the same small organization or division, (c) if the assessments were done at about the same time, or (d) if the criterion measures came from the same types of raters. Validities for different criteria or other variables under study were not combined. Key results for the original total sample and the supplementary, combined total sample are shown in Table 9. Differences are quite small, and we decided to proceed with analyses on the 107 validities.

Table 10 presents the variance of the weighted validities decomposed into their artifactual and nonartifactual components. Values of σ_e^2 depict the amount of variance due to sampling error (see Hunter et al., 1982, p. 44). Values of $\sigma^2\sqrt{r_{yy}} + \bar{U}$ represent the amounts of variance due to the combination of unreliability of the criterion and range restriction. These quantities are calculated using terms and their operations to the right of the minus sign in the numerator of the formula for $\sigma_{\rho_{xU}}^2$ (see Hunter et al., 1982, p. 90). Values of σ_1^2 represent the amount of variance remaining when σ_e^2 and $\sigma^2\sqrt{r_{yy}} + \bar{U}$ are removed. The final column contains the percentage of variance in the original validities that is not explained by statistical artifacts.

In the total sample and in five of the subgroups of validities, more than 40% of the variance in correlations could not be explained by artifacts. However, all of the variance for career progress, early identification, and research validities appears to be artifactual.

Table 8
Weighted Means and Variances Corrected for Artifacts

Sample	Weighted \bar{r}	Weighted s_r^2	$\bar{\rho}_{xU}$	$\sigma_{\rho_{xU}}^2$	\bar{c}	σ_c^2	$E(r_{yy}^{1/2})$	$\sigma^2 \sqrt{r_{yy}}$	Lower 90% credibility value	95% confidence interval	
										Lower bound	Upper bound
Total	.29	.0228	.37	.0171	.896	.032	.871	.015	.21	.11	.63
Criteria											
Performance	.25	.0233	.36	.0203	.902	.031	.774	.016	.18	.08	.64
Potential	.40	.0330	.53	.0373	.974	.004	.774	.016	.28	.15	.91
Dimension	.22	.0606	.33	.0998	.883	.028	.774	.016	-.07	-.29	.95
Training	.30	.0219	.35	.0197	.977	.004	.894	.002	.17	.07	.63
Career	.30	.0087	.36	.0000	.837	.051	1.000	.000	.36	.36	.36
Purpose											
Promotion	.24	.0304	.30	.0293	.939	.011	.871	.015	.08	-.04	.64
Early ID	.30	.0032	.46	.0000	.746	.056	.871	.015	.46	.46	.46
Selection	.29	.0166	.41	.0032	.805	.059	.871	.015	.34	.30	.52
Research	.42	.0027	.48	.0000	1.000	.000	.871	.015	.48	.48	.48

Note. Weighted \bar{r} = mean validity weighted by sample size; weighted s_r^2 = variance of validities weighted by sample size; $\bar{\rho}_{xU}$ = mean validity corrected for statistical artifacts; $\sigma_{\rho_{xU}}^2$ = variance corrected for statistical artifacts; \bar{c} = a measure of range restriction (1 = none, 0 = severe); σ_c^2 = variance of c; $E(r_{yy}^{1/2})$ = average of square roots of reliabilities across criterion measures; $\sigma^2 \sqrt{r_{yy}}$ = variance of $\sqrt{r_{yy}}$.
a When outliers were included, the total mean was .33, the career mean .34, and the promotion mean .32; all other means remained unchanged.
b The chi-square test of variance was significant ($p < .05$) for all but the career, early ID, selection, and research categories.

Moderator Analyses

Tables 11 and 12 contain the results of the moderator analyses. Table 11 presents the analyses for the two variables that had more than two categories: study design and publication form. The corrected means for all categories of both variables are similar. The corrected mean validities for different study designs ranged from .36 for experimental studies to .43 for predictive studies without feedback. The corrected mean validities for studies published in different forms ranged from .33 for presentations to .39 for unpublished technical reports. Thus, it appears that neither design of the study nor publication form moderate assessment center validity.

In one of our analyses we subdivided our total sample of validities into subgroups of validities based on both criterion type and assessment purpose and then attempted to assess differential moderating effects for study design and publication form within each of these subdivisions. We do not report these analyses here because in some subcategories there were too few validities (i.e., less than five) to ensure stability of the results. In other subcategories, we judged that insufficient true variance remained for a particular criterion or purpose category to permit

the operation of moderators. The latter judgment was made on the basis of the absolute amount of true variance found in the category, the results of the chi-square test on the uncorrected variances, and the percentage of the variance in the original correlations that could be explained by statistical artifacts. In a few subcategories, in which sufficient numbers of validities and variance existed, we found no support for differential moderating effects of study design or publication form within studies of different criteria and purpose. (These results may be obtained from the second author.)

Table 12 contains the results for analyses of potential moderators that are continuous and dichotomous variables. These variables were tested within the total sample of validities, within the job performance, potential, and dimension criterion categories, and within studies done for promotion purposes. Studies using career advancement criteria and studies conducted for selection, early identification, and research purposes were not analyzed because we judged that insufficient true variance existed (see Table 10).

The first row of entries for each potential moderator are Pearson product-moment and point-biserial correlations between the moderator variable and the effect size. The second row contains these correlations corrected for sampling error (Hunter et al., 1982, p. 52). The entries in parentheses are the number of validities used in the calculations. Due to the likelihood of capitalizing on chance with small samples, we excluded those correlations from the table that were based on fewer than nine validities. In sum, 69 correlations were computed and 25 were found significant ($p < .05$). The probability of this occurring by chance is extremely small ($CR = 14.77$, $p < .001$; Brozek & Tiede, 1952).

A few variables demonstrated significant correlations across samples of validities. The results suggest that assessment center

Table 9
Meta-Analytic Results on Total Sample: Before and After Combining Potentially Nonindependent Validities Within Studies

Sample	No. of validities	Weighted \bar{r}	Weighted s_r^2	$\bar{\rho}_{xU}$	$\sigma_{\rho_{xU}}^2$
Before combining	107	.2913	.02281	.3732	.01711
After combining	89	.2854	.02425	.3600	.01987

Table 10
Percentage of Weighted Variance Unexplained by Artifacts

Sample	Weighted s_r^2	σ_e^2	$\sigma^2\sqrt{r_{yy}} + \bar{U}$	σ_1^2	% variance unexplained by artifacts
Total	.0228	.0073	.0051	.0104	46
Criteria					
Performance	.0233	.0092	.0041	.0100	43
Potential	.0330	.0069	.0049	.0212	64
Dimension	.0606	.0109	.0031	.0466	77
Training	.0219	.0062	.0006	.0151	69
Career	.0087	.0055	.0067	.0000	0
Purpose					
Promotion	.0304	.0089	.0019	.0196	65
Early ID	.0032	.0060	.0107	.0000	0
Selection	.0166	.0058	.0093	.0015	9
Research	.0027	.0049	.0035	.0000	0

Note. Weighted s_r^2 = variance of validities weighted by sample size. σ_e^2 = variance due to sampling error; $\sigma^2\sqrt{r_{yy}} + \bar{U}$ = variance due to range restriction and unreliability on the criterion; σ_1^2 = variance left over after removing artifacts.

validities are higher when the percentage of male assesseees is low, when a larger number of assessment devices are used, when assessors are psychologists rather than managers, when peer evaluations are used, and when the studies are judged to be of higher quality.

Other variables, however, operated as moderators only within a group of studies that were conducted for a single purpose or that used a particular criterion. For example, within the group of studies done for promotion purposes, validities are higher when the percentage of minority assesseees is low. When predicting job performance, lower validities were found when assessors spend more days observing assesseees. In addition, when ratings of potential are the criterion, validities are higher when feedback is given to assesseees than when it is not.

Discussion

Generalizability of Assessment Centers

The findings of this meta-analysis support the widely held contention that assessment centers have predictive validity (By-

ham, 1970; Cohen et al., 1977; Howard, 1974; Huck, 1977; Hunter & Hunter, 1984; Thornton & Byham, 1982). Assessment center validities, corrected for sampling error, restriction of range, and criterion unreliability yielded a mean validity coefficient of .37. The average corrected validity coefficients for the various purposes of assessment centers ranged from .30 in promotional studies to .48 in basic research studies. Mean corrected validity coefficients for the prediction of different criteria ranged from .33 for dimensional ratings to .53 for ratings of management potential. Given the lower bound of the 90% credibility value for the average corrected validity coefficient in total sample, .21, we conclude that the validity of assessment centers does generalize.

The results of this meta-analysis must be interpreted with caution because of the complex and variable nature of assessment centers. Reliance on these results assumes that a new assessment center application will be designed and administered as well as or better than the average assessment center reviewed in this study. The *Standards and Ethical Considerations for Assessment Center Operations* (Task Force on Assessment Center

Table 11
Corrected Means and Variances for Study Design and Publication Form Calculated Across the Total Sample of Validities

Sample	No. of studies	No. of validities	Weighted \bar{r}	Weighted s_r^2	$\bar{\rho}_{xU}$	$\sigma_{\rho_{xU}}^2$	\bar{c}	σ_c^2
Study design								
Experiment	7	15	.32	.0189	.36	.0161	1.000	.000
Predictive (w/o feedback)	7	14	.30	.0311	.43	.0107	.809	.052
Predictive (w/feedback)	23	59	.29	.0234	.39	.0186	.855	.039
Concurrent	10	15	.36	.0184	.42	.0035	1.000	.000
Publication form								
Journal	25	58	.30	.0188	.38	.0110	.916	.030
Unpublished	10	21	.32	.0267	.39	.0194	.927	.021
Presentation	10	25	.23	.0272	.33	.0303	.812	.041

Note. Weighted \bar{r} = mean validity weighted by sample size; weighted s_r^2 = variance of validities weighted by sample size; $\bar{\rho}_{xU}$ = mean validity corrected for artifacts; $\sigma_{\rho_{xU}}^2$ = variance corrected for statistical artifacts; \bar{c} = a measure of range restriction (1 = none, 0 = severe); σ_c^2 = variance of c across validities.

Standards, 1980) provides guidance on the essential features of an assessment center.

Note that the present corrected validity coefficients differ from those calculated by Hunter and Hunter (1984). Hunter and Hunter found median corrected correlations of .63 for potential and .43 for performance, compared to the present mean correlations of .53 and .36, respectively. Considering that we corrected for sampling error, range restriction, and differences in unreliability in the criterion, whereas Hunter and Hunter (1984) corrected only for the first artifact, one would expect our values to be higher. Our lower values may be due to two factors: (a) We included a wider selection of studies, both published and unpublished, and additional studies conducted in the last 11 years, and (b) more recent studies tend to have lower validity as indicated by the slight negative correlation between publication year and assessment center validities. Taken together, the two meta-analyses suggest that assessment centers show validity generalization.

It should be recognized that the validity coefficients used for this meta-analysis may reflect a subtle form of *criterion contamination* not ferreted out in our moderator analyses of study design, study quality, and type of criterion (all of which are discussed later). We are referring to a set of perceptions about the qualities of a good manager that may be shared by the assessors (usually managers themselves) and anyone who provides criterion data later (e.g., performance ratings or promotion decisions). What we call a *validity coefficient* may be partially determined by a prototype (Feldman, 1981) of "a good manager" held in common among the various people providing both predictor and criterion data. This hypothesis deserves further investigation.

Situational Specificity of Assessment Centers

Our results also provide support for the situational specificity of assessment centers. Whereas recent validity generalization studies have shown that sampling error, unreliability of predictors and criteria, and range restriction account for about 75% of the observed variance across test validation studies (Hunter, 1980; Lilienthal & Pearlman, 1983; Pearlman, 1984; Pearlman et al., 1980; Schmidt et al., 1980; Schmidt & Hunter, 1977; Schmidt et al., 1981; Schmidt, Hunter, Pearlman, & Shane, 1979), these statistical artifacts accounted for only 54% of the observed variance of the total sample in the present study. Therefore, almost one half of the variance remains unexplained. Percentages of variance unexplained for studies conducted for certain purposes and involving some criteria were even higher. These results may be explained by the fact that the assessment center is a general method characterized by different procedures.

In addition to the percentage of remaining observed variance, it is important to note that the absolute level of variance is substantial. For the total sample, the standard deviation of true validities is .13, and for studies using dimension ratings it is .32. Even if artifacts for which we did not correct could account for one half of this remaining variance, there would still be enough variance remaining to conclude that assessment centers do show different true levels of validity. This finding is consistent with the conclusions of Schmidt and Hunter and their colleagues that validity generalization is frequently possible

even when the situational specificity hypothesis cannot be rejected (Pearlman et al., 1980, 1981; Schmidt et al., 1980). Support for both validity generalization and situational specificity has been found for weighted application blanks (Brown, 1981), intelligence and arithmetic tests (Schmidt et al., 1981), and the Law School Aptitude Test (Linn, Harnisch, & Dunbar, 1981). Given the utility work of Brown (1981), which suggests that even small *real* differences in validity coefficients across situations can have practical monetary implications, we suggest that designers of assessment centers take into consideration the variables found to moderate validities in this study.

Moderators of Assessment Center Validity

We looked for moderators within several coding categories: assessee characteristics, evaluation device characteristics, other assessment center characteristics, and validation study characteristics. These findings must be interpreted with caution because of the small sample sizes in some analyses.

Assessee characteristics. Assessee age, sex, and minority status were analyzed as potential moderators. There was no relation between average age of assessees and predictive validity of the assessment center. However, results suggest that assessment centers are more valid when the composition of the group consists of a larger proportion of women and a smaller proportion of minorities. Two explanations are possible: (a) Assessment centers may be more valid for women and for minorities, or (b) group composition alters the dynamics of the assessment process such that the overall assessment rating is more accurate when the assessee group includes a large portion of women, and less accurate when the assessee group includes a large portion of minorities.

The first explanation is not supported by previous studies of assessment centers, which have found no differential validity for Blacks and Whites (Huck, 1974; Huck & Bray, 1976), or for men and women (S. Alexander, 1975; Clingenpeel, 1979; Hall, 1976; Marquardt, 1976; Moses, 1973b; Moses & Boehm, 1975). Because women may be more self-disclosing than men (see, e.g., Fletcher, 1981; Fletcher & Spencer, 1984), they may provide assessors with more or better information to help make assessment ratings.

It is also possible to rule out other explanations for the present finding that sex and predictive validity of the assessment center are related, by examining correlations of sex and other moderators. For example, percentage of women is inversely related to the use of peer evaluations and the use of psychologists as assessors. Because studies that use peer evaluations and psychologists as assessors have higher validities, we can have greater confidence that sex is itself a moderator.

Evidence related to the second explanation, which is that group composition affects the dynamics of assessment, comes from a study by Schmitt and Hill (1977), who found that peer and assessor average ratings were minimally influenced by the proportion of men and women, or Blacks and Whites, in the group. The ratings of Black women on some dimensions were somewhat lower when the group consisted of a large portion of White men. Further study is needed to determine whether group composition or differential validity explains the findings of this meta-analysis.

It may initially appear unsettling that for promotional stud-

ies, the greater the proportion of minority assesseees, the lower the validity of the assessment center. However, in the spirit of affirmative action and in response to pressures from compliance agencies, organizations may be promoting greater numbers of minority candidates even though they have received relatively low assessment ratings.

Evaluation device characteristics. The analyses suggest that assessment centers are more predictively valid when a greater number of different types of exercises are used. One might discount this finding if the variable is positively correlated with other moderators. In fact, we found that the number of types of exercises is negatively correlated with the use of peer evaluations, which also moderates assessment center validities, and thus we can be somewhat more confident that it actually moderates validities. This supports the advice (Thornton & Byham, 1982) that a broad spectrum of types of exercises should be used to attain content representativeness in assessment centers.

Although less than one half of the predictive validity studies reviewed used some form of peer evaluation to help evaluate assesseees, assessment centers that did were found to be more valid than those that did not. This finding is not surprising given the substantial amount of evidence that shows that peer assessment can be both a reliable and valid predictor of performance (Kane & Lawler, 1978). Although organizations are reluctant to use peer ratings for fear of increasing competitiveness among assesseees, they should be used more often in the future to supplement the ratings of trained assessors if such reactions could be minimized.

Other assessment center characteristics. We also analyzed a number of other assessment center characteristics to see whether they moderated assessment center validity. These were, type of assessor used (i.e., manager or psychologist), amount of training assessors were given, the number of days assessors spent observing assesseees and the number of hours they spent integrating information, the ratio of the number of assesseees to assessors, and whether feedback was given to assesseees or to their immediate supervisors.

Type of assessor was the only variable among these assessment center characteristics that moderated validities in the total sample. In contrast to other researchers (Greenwood & McNamara, 1969; Thomson, 1970) who have found no difference in the assessment center ratings of professional (i.e., psychologists) and nonprofessional (i.e., in-house managers) assessors, we found evidence that assessment centers that use psychologists as assessors are significantly more valid than those that use managers as assessors. Many people in the field believe that managers are better able to interpret the meaning of different behaviors for a particular job than are psychologists, because they are more familiar with the requirements of the job. However, the results of this meta-analysis suggest that psychologists provide more valid assessment center ratings than do managers. In fact, this moderator is particularly robust given that it is negatively related to many other moderators.

The following variables, all thought to be related to assessment center validity, were not significantly correlated with validities in this meta-analysis: ratio of assesseees to assessors, number of days of observation and days of assessor training, hours spent integrating information, feedback to assesseees and their supervisors, and criterion contamination. Providing feedback to assesseees and their supervisors seems to have little effect

on assessment center validities. (Only when potential ratings are the criterion does feedback to assesseees seem to inflate validities.) These findings suggest that criterion contamination is not the sole explanation for the high correlation of assessment center and follow-up ratings.

One finding that initially surprised us was that amount of assessor training did not affect the validity of assessment centers. Thorough assessor training is thought to be essential to producing reliable and valid ratings (Task Force on Assessment Center Standards, 1980). However, given the mixed success of assessor training for related skills (see Landy & Farr, 1980), the lack of relation among assessor training and assessment center validity found in this meta-analysis is not very surprising.

The present results should be interpreted with caution, however, because research reports do not always give adequate descriptions of the amount or type of assessor training. Although there were no reports of research that did *not* train assessors, a number of researchers failed to mention whether assessors were trained and if they were, for how long. Therefore, we were unable to discern whether there is a significant difference in the validity of ratings of assessors who have been trained compared to those who have not. However, we can conclude that within the range of number of days of training studied (.5–15), more training does not lead to high validities.

Validation study characteristics. The results of this meta-analysis support those who maintain that assessment centers are more valid for predicting an assessee's job potential ($\bar{\rho} = .53$) than for predicting performance ($\bar{\rho} = .36$).

These analyses are quite comparable to the analyses conducted by Cohen et al. (1977). Cohen and his colleagues concluded that predictive accuracy was highest for job potential ($Mdn r = .63$), followed by progress ($Mdn r = .40$) and job performance ($Mdn r = .33$). In addition, subsequent individual studies by Klimoski and Strickland (1981) and Turnage and Muchinsky (1984) found that assessment centers predict progress but not performance criteria.

As Klimoski and Strickland (1977) pointed out, the superior ability of assessment centers to predict potential over performance may be due to the assessment staff's intuitive grasp of organizational values and norms with regard to promotion, and to their adeptness at predicting who will get promoted in the organization. Predicting an assessee's subsequent job performance, given the variety of factors outside the assessee's immediate control (e.g., dependency on other workers, customers, raw materials) and the notorious bias of supervisory ratings is a much more challenging task.

Another validation study characteristic we investigated was study design (i.e., whether the validity study was a predictive study with or without feedback, a concurrent validation, or a pure experiment). Our results show that study design does not moderate assessment center validities, a finding that supports research on cognitive tests (Bemis, 1968; Pearlman et al., 1980). This finding, along with the lack of significant correlation between validities and potential for criterion contamination through knowledge of the assessment results, contradicts a popular belief that operational use of assessment center data inflates validity coefficients as a result of contamination via knowledge of the predictor data. If contamination was a serious problem, validities for studies that operationally used assessment center data would be much higher. Our meta-analysis found no sig-

Table 12
Moderators for the Total Sample and Selected Criterion and Purpose Categories

Moderator	Total sample	Criteria type			Purpose of assessment center: Promotion
		Job performance	Potential ratings	Dimension ratings	
Publication year					
<i>r</i>	-.13	-.08	.11	-.87*	-.52*
ρ^a	-.16	-.10	.13	-.96	-.62
No. of validities	107	44	13	9	51
Mean age of assessees					
<i>r</i>	.06	.15	-.51	—	-.19
ρ^a	.07	.26	-.58	—	-.23
No. of validities	57	19	10	—	23
Percentage men					
<i>r</i>	-.43*	-.55*	-.91*	—	-.26
ρ^a	-.52	-.71	-1.00	—	-.31
No. of validities	68	28	9	—	31
Percentage minority					
<i>r</i>	.03	.01	—	—	-.74*
ρ^a	.03	.02	—	—	-.88
No. of validities	37	15	—	—	19
Used a general mental ability test					
<i>r</i>	-.11	.17	-.24	-.91*	-.06
ρ^a	-.14	.22	-.26	-1.00	-.07
No. of validities	106	43	12	9	49
Total number of devices					
<i>r</i>	.25*	.19	.63*	.86*	.48*
ρ^a	.31	.25	.71	.95	.57
No. of validities	104	42	12	9	47
Ratio of assessors to assessees					
<i>r</i>	-.12	-.14	-.26	—	-.17
ρ^a	-.15	-.18	-.29	—	-.20
No. of validities	80	33	11	—	43
Days of observation					
<i>r</i>	-.02	-.50*	.14	—	.03
ρ^a	-.02	-.64	.16	—	.03
No. of validities	96	37	12	—	42
Days of assessor training					
<i>r</i>	.08	.00	-.35	.17	-.22
ρ^a	.10	.01	-.39	.18	-.26
No. of validities	67	28	10	9	42
No. of hours spent integrating information					
<i>r</i>	-.01	-.33	-.22	—	-.19
ρ^a	-.02	-.42	-.25	—	-.23
No. of validities	53	19	11	—	36
Psychologist vs. managers as assessors ^b					
<i>r</i>	-.21*	NV	-.34	—	-.29*
ρ^a	-.26	NV	-.39	—	-.34
No. of validities	76	31	11	—	44
Peer evaluation ^c					
<i>r</i>	.36*	.20	.62*	.91*	.28*
ρ^a	.44	.26	.70	1.00	.33
No. of validities	93	40	12	9	47
Feedback given to assessees ^d					
<i>r</i>	.10	.07	.62*	—	.18
ρ^a	.12	.09	.70	—	.21
No. of validities	87	31	12	—	39
Feedback given to immediate supervisor ^d					
<i>r</i>	-.14	-.17	-.15	—	-.03
ρ^a	-.17	-.22	-.58	—	-.03
No. of validities	77	25	12	—	29

Table 12 (continued)

Moderator	Total sample	Criteria type			Purpose of assessment center: Promotion
		Job performance	Potential ratings	Dimension ratings	
Criterion contamination					
<i>r</i>	-.07	-.24	-.18	-.17	-.17
<i>rho</i> ^a	-.08	-.32	-.20	-.18	-.20
No. of validities	105	43	12	9	49
Quality of study (summed rating)					
<i>r</i>	.15†	-.18	.66*	.90*	.14
<i>rho</i> ^a	.18	-.23	.74	.99	.17
No. of validities	105	43	12	9	49
Quality of study (overall rating)					
<i>r</i>	.26*	.23	.21	.91*	.33*
<i>rho</i> ^a	.31	.29	.24	1.00	.39
No. of validities	107	44	13	9	51

Note. Significance tests were not performed on the rhos because no such test exists. NV = no variance in the moderator variable. All assessors were managers in the studies in which performance was the criterion. Descriptive data in each category can be obtained from the second author.
^a Corrected for sampling error. ^b Psychologists coded 1; managers coded 2. ^c Absence of peer evaluation coded 1; presence of peer evaluation coded 2. ^d Feedback not given coded 1; feedback given coded 2.
* *p* < .05. † *p* < .06.

nificant differences between studies that operationally used assessment center data and those that did not. In combination, the findings refute the contention that direct contamination explains the observed validities of assessment centers.

A major caveat pervading our own analyses has been the internal and external validity of the studies. We found that the degree to which validation studies are internally and externally valid is related to their predictive validity. Our rating of the quality of the study, based on the representativeness of the sample, and motivational, job experience, and training differences between assesseees and present employees was highly correlated with the validity of the assessment center. This finding supports Thornton and Byham's (1982) observation that methodologically sound studies have higher validity.

One somewhat unexpected finding is the lack of a significant relation between assessment center validities and the time at which criterion measures are taken. Much of the prior research found overall assessment ratings to be more predictive over a longer period of time (Hinrichs, 1978; Mitchel, 1975; Moses, 1972). Other researchers have found no relation between validities and time of criterion measure (Finley, 1970) or have found support for a negative relation (Howard, 1979; Slivinski & Bourgeois, 1977). Clearly this issue deserves further research.

In conclusion, we recommend that assessment centers be designed to use the features that are associated with the more highly valid programs reviewed in this meta-analysis. A well-designed assessment center will probably have predictive validity, but to optimize validity, certain procedures should be followed. Our results suggest that in the future, assessment centers should include more assessment devices, use psychologists as assessors, and supplement assessor ratings with those provided by peers. Within the range of variables reviewed in this meta-analysis, it does not appear that there is a systematic relation between the size of assessment center validities and the length of assessor training, time lapsed between the assessment center

and when criterion measures are taken, the number of hours assessors spend integrating information, the number of days assesseees are observed, or whether assessment center data is operationally used, if precautions are taken. We also recommend that validation studies of assessment centers be conducted with adequate research methodology (e.g., ensuring adequate sample representativeness).

The Art and Science of Meta-Analysis

After completing this meta-analysis, we believe that conducting a validity generalization study is somewhat of an art. Judgments are required at many junctures. Some of the issues we found most difficult to resolve are discussed ahead.

Moderator analyses. We studied moderator variables in a number of ways. First, studies were presorted into groups when the variables were categorical and there were a priori reasons for doing so. Then meta-analyses were performed in each group. This was the approach taken with type of criterion and assessment purpose. Both variables have been handled this way in prior meta-analytic work. This approach seems appropriate when the categories such as criterion types are theoretically and logically distinct from one another. However, type of criterion can also be viewed as just one of many assessment center design variables that vary from study to study and therefore should be treated as any other potential moderator. Hence, potential moderators should be tested only within the total sample of studies. We decided to test for moderators both within the total sample and within subgroups of studies using the same type of criterion.

In theory, our research allows us to compare the results of searching for moderators within the total sample and within presorted categories. Unfortunately, we were unable to completely carry out this comparison. Although we began our analyses with 107 validity coefficients, after presorting studies by

criterion and purpose, we quickly reached the point at which there were not enough studies in some categories to make meaningful comparisons on some variables. This may be more of a problem for assessment centers than other selection devices because of their complex design. We encourage others to study this issue further, using a selection device for which a larger number of studies exist.

A second method used for analyzing moderators was to compute point-biserial correlations between dichotomous moderators and validity coefficients. If a strong relation between a variable and validity is not found, it is highly unlikely that mean correlations will vary from subset to subset. The advantage of this approach is that it requires fewer calculations than performing meta-analyses within each subset of studies. Thus, point-biserial correlations can potentially be a valuable time-saving strategy. For continuous variables, we correlated the moderator with the validities using the Pearson product-moment method.

A problem encountered when interpreting the results of the moderator analyses was the interdependency of many of the variables. One way of handling correlated moderators is to partial out the effects due to one moderator and then look at the correlation between another variable and the mean validity. Unfortunately, sample sizes were too small to enable us to meaningfully do this. Instead, we were forced to speculate about how the interdependency among variables affects the strength of the moderators individually.

Large-sample studies. The validity generalization literature is not clear about how to decide whether to exclude studies with unusually large samples. We chose to exclude three studies because we did not want them to have undue influence on the mean and variance estimates. Yet, it can be argued that large studies have little sampling error; therefore, their influence is legitimate. In the present case, comparisons showed that the results are the same regardless of whether those outliers are included or not!

Insufficient reporting. Finally, we echo Schmidt et al.'s (1980) and Orwin and Cordray's (1985) contention that reports of validity studies must be more complete for validity generalization and meta-analysis research to be maximally effective. Orwin and Cordray found that deficient reporting injects considerable noise into meta-analysis data that can lead to spurious conclusions. Although we heeded their recommendations for counteracting the effects of deficient reporting (i.e., computed separate reliability coefficients for individual coding items, based on appropriate estimators; incorporated data quality information into our analysis; obtained additional information on primary studies by contacting original investigators), we were still unable to code all of the study features from most reports. In fact, a few studies were totally eliminated from the analyses because they failed to report enough essential information. In particular, it was difficult to evaluate the amount of range restriction present in assessment center validity studies and to estimate the reliability of criteria. We had to obtain surrogate data from other related areas of literature to construct some distributions. For example, we also used reliabilities from the performance evaluation literature to help construct reliability distributions of assessment center criterion ratings.

Deficits in reporting reinforce our notion that validity generalization is still somewhat of an art. The general procedure is

well laid out, but there are numerous stages in the analysis in which judgment comes into play. How to combine most meaningfully across effect sizes, deal with the problems of unusually large sample sizes and correlated moderators, and obtain surrogate estimates to construct distributions, are but a few.

References

- Alexander, L. D. (1979). An exploratory study of the utilization of assessment center results. *Academy of Management Journal*, 22, 152-157.
- Alexander, S. J. (1975). Bendix Corporation establishes early identification program. *Assessment and Development*, 2, 10.
- Bemis, S. E. (1968). Occupational validity of the General Aptitude Test Battery. *Journal of Applied Psychology*, 52, 240-244.
- Bender, J. M. (1973). What is "typical" of assessment centers? *Personnel*, 50, 50-57.
- Bray, D. W., & Grant, D. L. (1966). The assessment center in the measurement of potential for business management. *Psychological Monographs*, 80 (17, Whole No. 625).
- Brown, S. H. (1981). Validity generalization in the life insurance industry. *Journal of Applied Psychology*, 66, 664-670.
- Brozek, J. & Tiede, K. (1952). Reliable and questionable significance in a series of statistical tests. *Psychological Bulletin*, 49, 339-341.
- Burroughs, W. A., Rollins, J. B., & Hopkins, J. J. (1983). The effects of age, departmental experience, and prior rater experience on performance in assessment center exercises. *Academy of Management Journal*, 16, 335-339.
- Byham, W. C. (1970). Assessment center for spotting future managers. *Harvard Business Review*, 48, 150-160.
- Byham, W. C. (1978a). How to improve the validity of an assessment center. *Training and Development Journal*, 32, 4-6.
- Byham, W. C. (1978b, July). *Intercultural adaptability of the assessment center method*. Paper presented at Nineteenth International Congress of Applied Psychology, Munich, FRG.
- Byham, W. C. (1981). *Dimensions of managerial success*. Pittsburgh, PA: Development Dimensions International.
- Clingenpeel, R. (1979, June). *Validity and dynamics of a foreman selection process*. Paper presented at the meeting of the Seventh International Congress on the Assessment Center Method, New Orleans.
- Cohen, B. M., Moses, J. L., & Byham, W. C. (1977). *The validity of assessment centers: A literature review* (Rev. ed.; Monograph No. 2). Pittsburgh, PA: Development Dimensions Press.
- Cook, T. D., & Campbell, D. T. (1976). The design and conduct of quasi-experiments and true experiments in field settings. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 223-326). Chicago: Rand McNally.
- Cooper, H. M., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin*, 87, 442-449.
- Feldman, J. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology*, 66, 127-148.
- Finley, R. M., Jr. (1970, September). *An evaluation of behavior predictions from projective tests given in a management assessment center*. Paper presented at the 78th Annual Convention of the American Psychological Association, Miami Beach.
- Fletcher, C. (1981). The influence of candidates' beliefs and self-presentation strategies in selection interviews. *Personnel Review*, 10, 14-17.
- Fletcher, C., & Spencer, A. (1984). Sex of candidate and sex of interviewer as determinants of self-presentation orientation in interviews: An experimental study. *International Review of Applied Psychology*, 33, 305-313.
- Greenwood, J. M., & McNamara, W. J. (1969). Leadership styles of structure and consideration and managerial effectiveness. *Personnel Psychology*, 22, 141-152.

- Hall, H. L. (1976). *An evaluation of the upward mobility assessment center for the Bureau of Engraving and Printing* (TM No. 76-6). Washington, DC: U.S. Civil Service Commission.
- Hinrichs, J. R. (1978). An eight-year follow-up of a management assessment center. *Journal of Applied Psychology*, 63, 596-601.
- Howard, A. (1974). An assessment of assessment centers. *Academy of Management Journal*, 17, 115-134.
- Howard, A. (1979, June). *Assessment center predictions sixteen years later*. Paper presented at the meeting of the Seventh International Congress on the Assessment Center Method, New Orleans.
- Huck, J. R. (1973). Assessment centers: A review of the external and internal validities. *Personnel Psychology*, 26, 191-212.
- Huck, J. R. (1974). *Determinants of assessment center ratings for White and Black females and the relationship of these dimensions to subsequent performance effectiveness*. Unpublished doctoral dissertation, Wayne State University.
- Huck, J. R. (1977). The research base. In J. L. Moses & W. C. Byham (Eds.), *Applying the assessment center method* (pp. 261-291). New York: Pergamon Press.
- Huck, J. R., & Bray, D. W. (1976). Management assessment center evaluations and subsequent job performance of Black and White females. *Personnel Psychology*, 29, 13-30.
- Hunter, J. E. (1980). *Test validation for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB)*. Unpublished manuscript, Michigan State University.
- Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, 96, 72-98.
- Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Advanced meta-analysis: Quantitative methods for cumulating research findings across studies*. Beverly Hills, CA: Sage.
- Jaffee, C. L., Cohen, S. L., & Cherry, R. (1972). Supervisory selection program for disadvantaged or minority employees. *Training and Development Journal*, 26, 22-28.
- Kane, J. S., & Lawler, E. E., III. (1978). Methods of peer assessment. *Psychological Bulletin*, 85, 555-586.
- Klimoski, R. J., & Strickland, W. J. (1977). Assessment centers: Valid or merely prescient. *Personnel Psychology*, 30, 353-363.
- Klimoski, R. J., & Strickland, W. J. (1981). *A comparative view of assessment centers*. Unpublished manuscript.
- Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, 87, 72-107.
- Lilienthal, R. A., & Pearlman, K. (1983). *The validity of Federal selection tests for aide technicians in the health, science, and engineering fields* (OPRD Report No. 83-1). Washington, DC: U.S. Office of Personnel Management, Office of Personnel Research and Development. (NTIS No. PB83-202051)
- Linn, R. L., Harnisch, D. L., & Dunbar, S. B. (1981). Validity generalization and situational specificity: An analysis of the prediction of first year grades in law school. *Applied Psychological Measurement*, 5, 281-289.
- Marquardt, L. D. (1976). *Follow-up evaluation of the second look approach to the selection of management trainees*. Chicago: Sears, Roebuck and Company, National Personnel Department, Psychological Research and Services.
- Mitchel, J. O. (1975). Assessment center validity: A longitudinal study. *Journal of Applied Psychology*, 60, 573-579.
- Moses, J. L. (1972). Assessment center performance and management progress. *Studies in Personnel Psychology*, 4, 7-12.
- Moses, J. L. (1973a). Assessment center for the early identification of supervisory and technical potential. In W. C. Byham & D. Bobin (Eds.), *Alternatives to paper and pencil testing* (pp. 38-49). Pittsburgh, PA: University of Pittsburgh. [Proceedings of a conference at Graduate School of Business]
- Moses, J. L. (1973b). The development of an assessment center for the early identification of supervisory potential. *Personnel Psychology*, 26, 569-580.
- Moses, J. L., & Boehm, V. R. (1975). Relationship of assessment center performance to management progress of women. *Journal of Applied Psychology*, 60, 527-529.
- Neidig, R. D., Martin, J. C., & Yates, R. E. (1978). *The FBI's Management Aptitude Program Assessment Center* (Research Rep. No. 1, TM 78-3). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center, Applied Psychology Section.
- Orwin, R. G., & Cordray, D. S. (1985). Effects of deficient reporting on meta-analysis: A conceptual framework and reanalysis. *Psychological Bulletin*, 97, 134-147.
- Pearlman, K. (1984, August). *Validity generalization: Methodological and substantive implications for meta-analytic research*. Paper presented at the 92nd Annual convention of the American Psychological Association, Toronto, Ontario, Canada.
- Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict training success and job proficiency in clerical occupations. *Journal of Applied Psychology*, 65, 373-406.
- Ritchie, R. J., & Moses, J. L. (1983). Assessment center correlates of women's advancement into middle management: A 7-year longitudinal analysis. *Journal of Applied Psychology*, 68, 227-231.
- Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin*, 85, 185-193.
- Russell, G. (1975). Differences in minority/nonminority assessment center ratings. *Assessment and Development*, 3, 3, 7, 8.
- Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J. E. (1980). Validity generalization results for computer programmers. *Journal of Applied Psychology*, 65, 643-661.
- Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, 62, 529-540.
- Schmidt, F. L., Hunter, J. E., & Caplan, J. R. (1981). Validity generalization results for jobs in the petroleum industry. *Journal of Applied Psychology*, 66, 261-273.
- Schmidt, F. L., Hunter, J. E., Pearlman, K., & Shane, G. S. (1979). Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. *Personnel Psychology*, 32, 257-281.
- Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology*, 37, 407-422.
- Schmitt, N., & Hill, T. E. (1977). Sex and race composition of assessment center groups as a determinant of peer and assessor ratings. *Journal of Applied Psychology*, 62, 261-264.
- Slivinski, L. W., & Bourgeois, R. P. (1977). Feedback of assessment center results. In J. L. Moses & W. C. Byham (Eds.), *Applying the assessment center method* (pp. 143-159). New York: Pergamon Press.
- Task Force on Assessment Center Standards. (1980). Standards and ethical considerations for assessment center operations. *The Personnel Administrator*, 25, 35-38.
- Thomson, H. A. (1970). Comparison of predictor and criterion judgments of managerial performance using the multitrait-multimethod approach. *Journal of Applied Psychology*, 54, 496-502.
- Thornton, G. C. III, & Byham, W. C. (1982). *Assessment centers and managerial performance*. New York: Academic Press.
- Turnage, J. J., & Muchinsky, P. M. (1984). A comparison of the predictive validity of assessment center evaluations versus traditional measures in forecasting supervisory job performance: Interpretive implications of criterion distortion for the assessment paradigm. *Journal of Applied Psychology*, 69, 595-602.

Appendix

Studies Included in Meta-Analysis

- Alexander, H. S., Buck, J. A., & McCarthy, R. J. (1975). Usefulness of the assessment center process for selection to upward mobility programs. *Human Resource Management, 14*, 10–13.
- American Airlines. (1976). *A preliminary report on the validity of the Key Manager Human Resources Center*. New York: American Airlines, Personnel Resources Department.
- Anstey, E. (1966). The Civil Service Administrative Class and the Diplomatic Service: A follow-up. *Occupational Psychology, 40*, 139–151.
- Anstey, E. (1971). The Civil Service Administrative Class: A follow-up of post-war entrants. *Occupational Psychology, 45*, 27–43.
- Anstey, E. (1976). Civil Service administrators: A long-term follow-up. *Behavioral Sciences Research Division* (Report No. 31). London: HMS, Civil Service Department.
- Bentz, V. J. (1980, June). *Overview of Sears research with multiple assessment techniques*. Paper presented at the meeting of the Eighth International Congress on the Assessment Center Method, Toronto, Canada.
- Borman, W. C. (1982). Validity of behavioral assessment for predicting military recruits performance. *Journal of Applied Psychology, 67*, 3–9.
- Bray, D. W. (1964). The assessment center method of appraising management potential. In J. W. Blood (Ed.), *The personnel job in a changing world* (pp. 225–234). New York: American Management Association.
- Bray, D. W. (1982). The assessment center and the study of lives. *American Psychologist, 37*, 180–189.
- Bray, D. W., & Campbell, R. J. (1968). Selection of salesmen by means of an assessment center. *Journal of Applied Psychology, 52*, 36–41.
- Bray, D. W., Campbell, R. J., & Grant, D. L. (1974). *Formative years in business: A long-term AT&T study of managerial lives*. New York: Wiley.
- Bray, D. W., & Grant, D. L. (1966). The assessment center in the measurement of potential for business management. *Psychological Monographs, 80* (17, Whole No. 625).
- Bullard, J. F. (1969). *An evaluation of the assessment center approach to selecting supervisors*. Peoria, IL: Caterpillar Tractor Company.
- Campbell, R. J., & Bray, D. W. (1967). Assessment centers: An aid in management selection. *Personnel Administration, 30*, 6–13.
- Carleton, F. O. (1970, September). *Relationships between follow-up evaluations and information developed in a management assessment center*. Paper presented at the 78th Annual Convention of the American Psychological Association, Miami Beach.
- Dunne, G. J. Jr., Komar, D. M., Wise, W. W., & Norton, S. D. (1981, April). *An empirical look at an assessment center for R&D managers*. Paper presented at the Ninth International Congress on the Assessment Center Method, San Diego, CA.
- Erpenbach, J. J. (March 11, 1971, personal communication)
- Gardner, K. E., & Williams, A. P. O. (1973). A twenty-five year follow-up of an extended interview selection procedure in the Royal Navy. *Occupational Psychology, 47*, 1–13.
- Grossner, C. (1974). *The assessment of the assessment center*. Unpublished doctoral dissertation, Sir George Williams University, Montreal, Quebec, Canada.
- Haynes, M. E. (1978). *Operations supervisor assessment program: Five-year post program evaluation*. Unpublished report, Shell Oil Corporation, Houston, TX.
- Hinrichs, J. R. (1969). Comparison of “real life” assessment of management potential with situation exercises, paper-and-pencil ability tests, and personality inventories. *Journal of Applied Psychology, 53*, 425–432.
- Hinrichs, J. R. (1978). An eight-year follow-up of a management assessment center. *Journal of Applied Psychology, 63*, 596–601.
- Howard, A. (1979, June). *Assessment center predictions sixteen years later*. Paper presented at the Seventh International Congress on the Assessment Center Method, New Orleans.
- Huck, J. R., & Bray, D. W. (1976). Management assessment center evaluations and subsequent job performance of Black and White females. *Personnel Psychology, 29*, 13–30.
- Klimoski, R. J., & Strickland, W. J. (1981). *A comparative view of assessment centers*. Unpublished manuscript.
- Kraut, A. I., & Scott, G. J. (1972). Validity of an operational management assessment program. *Journal of Applied Psychology, 56*, 124–129.
- McConnell, J. J., & Parker, T. (1972). An assessment center program for multiorganizational use. *Training and Development Journal, 26*(3), 6–14.
- McElroy, J. J. (1979). *MacSteel Assessment Centers: Evaluation and validation report* (1977–1979). Unpublished manuscript.
- Metropolitan Transit Authority. (1972). *The uses of the assessment center in a government agency's management development program*. Unpublished report.
- Mitchel, J. O. (1975). Assessment center validity: A longitudinal study. *Journal of Applied Psychology, 60*, 573–579.
- Moses, J. L. (1972). Assessment center performance and management progress. *Studies in Personnel Psychology, 4*, 7–12.
- Moses, J. L., & Boehm, V. R. (1975). Relationship of assessment center performance to management progress of women. *Journal of Applied Psychology, 60*, 527–529.
- Moses, J. L., & Wall, S. (1975). Pre-hire assessment: A validity study a new approach for hiring college graduates. *Assessment and Development, 2*(2), 11.
- Norton, S. (1980, June). *Applying the assessment center method to an upward mobility program*. Paper presented at the meeting of the Eighth International Congress on Assessment Center Method, Toronto, Ontario, Canada.
- Parker, T. C. (1980). Assessment centers: A statistical study. *The Personnel Administrator, 25*, 65–67.
- Ritchie, R. J. (1980, June). *The validity of an assessment center for selecting telephone directory salespeople*. Paper presented at the meeting of the Eighth International Congress on the Assessment Center Method, Toronto, Ontario, Canada.
- Ritchie, R. J., & Moses, J. L. (1983). Assessment center correlates of women's advancement into middle management: A 7-year longitudinal analysis. *Journal of Applied Psychology, 68*, 227–231.
- Schmitt, N., Noe, R. A., Meritt, R., & Fitzgerald, M. P. (1984). Validity of assessment center ratings for the prediction of performance ratings and school climate of school administrators. *Journal of Applied Psychology, 69*, 207–213.
- Slivinski, L. W., Grant, K. W., Bourgeois, R. P., & Pederson, L. D. (1977). *Development and application of a first level management assessment centre*. Ottawa, Ontario, Canada: Personnel Psychology Centre, Managerial Assessment and Research Division.
- Slivinski, L. W., McCloskey, J. L., & Bourgeois, R. P. (1979, June). *Comparison of different methods of assessment*. Paper presented at the meeting of the Seventh International Congress on the Assessment Center Method, New Orleans.
- Thomson, H. A. (1969). *Internal and external validation of an industrial assessment program*. Unpublished doctoral dissertation, Case Western Reserve University.
- Tziner, A. (1982). *The assessment center goes to boot camp again: An*

- application to selection of officer training applicants.* Unpublished manuscript.
- Tziner, A. (1984). Prediction of peer rating in a military assessment center: A longitudinal follow-up. *Canadian Journal of Administrative Sciences*, 1, 146-160.
- Tziner, A., & Dolan, S. (1982). Validity of an assessment center for identifying future female officers in the military. *Journal of Applied Psychology*, 67, 728-736.
- Vernon, P. E. (1950). The validation of Civil Service Selection Board procedures. *Occupational Psychology*, 24, 75-95.
- Warriner, L. (1981, April). *Statistical vs. judgmental prediction of advancement using assessment center data.* Paper presented at the meeting of the Ninth International Congress on the Assessment Center Method, San Diego, CA.
- Wilson, N. A. (1948). The work of the Civil Service Selection Board. *Occupational Psychology*, 22, 204-212.
- Wissman, D. J., & Rankin, K. K. (1982). *A second look at the validity of a public sector assessment center for research and development managers.* Unpublished manuscript.
- Wollowick, H. B., & McNamara, W. J. (1969). Relationship of the components of an assessment center to management success. *Journal of Applied Psychology*, 53, 348-352.
- Worbois, G. M. (1975). Validation of externally developed assessment procedures for identification of supervisory potential. *Personnel Psychology*, 28, 77-91.

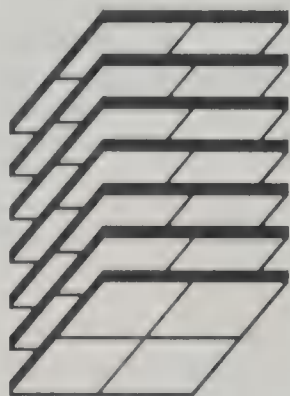
Received May 14, 1986

Revision received October 27, 1986

Accepted August 22, 1986 ■

*N*ow available . . .

A major revision of the



CALIFORNIA PSYCHOLOGICAL INVENTORY

by Harrison G. Gough

Featuring:

- A test booklet with 18 fewer items
- New, expanded edition of the **CPI Manual**
- Two new scales on the profile — **empathy** and **independence**
- Three new **structural scales** — role, character, and realization
- Extensive **empirically based narrative** report developed by the author
- **Microcomputer software** for administering, scoring, and reporting of the revised **CPI**

For complete information on the CPI revision write to:

ψ Consulting Psychologists Press
Box CPI-R
577 College Avenue
Palo Alto, CA 94306

Marygrove College Library
Detroit, Michigan 48221
PLEASE DO NOT REMOVE

Volume 72
Number 4

November 1987

Published quarterly
by the
American Psychological
Association

Journal of Applied Psychology

Editor

Robert M. Gulon

Associate Editors

Irwin L. Goldstein

Frank J. Landy

The *Journal of Applied Psychology* is devoted primarily to original investigations that contribute new knowledge and understanding to any field of applied psychology except clinical psychology. The journal considers quantitative investigations of interest to psychologists doing research or working in such settings as universities, industry, government, urban affairs, police and correctional systems, health and educational institutions, transportation and defense systems, and consumer affairs. A theoretical or review article may be accepted if it represents a special contribution to an applied field.

Editor

Robert M. Guion, *Bowling Green State University*

Associate Editors

Irwin L. Goldstein, *University of Maryland*

Frank J. Landy, *Pennsylvania State University*

Consulting Editors

Lewis E. Albright, *deRecat & Associates, San Francisco, California*

Earl A. Alluisi, *OUSDR, The Pentagon, Washington, DC*

Kenneth M. Alvares, *Frito-Lay, Dallas, Texas*

Phipps Arabie, *University of Illinois*

William B. Askren, *Air Force Human Resources Laboratory, Wright-Patterson Air Force Base, Ohio*

Kathryn M. Bartol, *University of Maryland*

Bernard M. Bass, *State University of New York, Binghamton*

Robert S. Billings, *Ohio State University*

Philip Bobko, *University of Kentucky*

C. Alan Boneau, *George Mason University*

Walter C. Borman, *Personnel Decisions Research Institute, Minneapolis, Minnesota*

Donald E. Broadbent, *University of Oxford, England*

Wayne F. Cascio, *University of Colorado, Denver*

Margaret M. Clifford, *University of Iowa*

H. Peter Dachler, *Hochschule St. Gallen für Wirts & Sozialwissen, St. Gallen, Switzerland*

Dan R. Dalton, *Indiana University*

Mark L. Davison, *University of Minnesota*

Robyn M. Dawes, *Carnegie-Mellon University*

Fritz Drasgow, *University of Illinois, Champaign*

Beverly Dugan, *New York Telephone, New York, New York*

E. Ralph Dusek, *JIL Systems and Services, Arlington, Virginia*

James L. Farr, *Pennsylvania State University*

Jack M. Feldman, *Georgia Institute of Technology*

Jeffrey H. Greenhaus, *Drexel University*

Tove Helland Hammer, *Cornell University*

William C. Howell, *Rice University*

Daniel R. Ilgen, *Michigan State University*

Andrew S. Imada, *University of Southern California*

Lawrence R. James, *Georgia Institute of Technology*

Stanislav V. Kasl, *Yale University*

James G. Kelly, *University of Illinois, Chicago*

Gary P. Latham, *University of Washington*

Edwin A. Locke, *University of Maryland*

Robert P. Lowman, *Kansas State University*

Ben B. Morgan, Jr., *University of Central Florida*

Karlene H. Roberts, *University of California, Berkeley*

Paul R. Sackett, *University of Illinois, Chicago*

Steven L. Sauter, *NIOSH, Cincinnati, Ohio*

Frank L. Schmidt, *University of Iowa*

Neal Schmitt, *Michigan State University*

Lyle F. Schoenfeldt, *Texas A&M University*

Stanley E. Seashore, *University of Michigan*

Kirk H. Smith, *Bowling Green State University*

Patricia Cain Smith, *Bowling Green State University*

Barry M. Staw, *University of California, Berkeley*

Mary L. Tenopir, *American Telephone & Telegraph Company, New York, New York*

James R. Terborg, *University of Oregon*

Gary L. Wells, *University of Alberta*

Gary A. Yukl, *State University of New York, Albany*

Sheldon Zedeck, *University of California, Berkeley*

Manuscripts: Submit manuscripts in quadruplicate to the Editor, Robert Guion, Department of Psychology, Bowling Green State University, Bowling Green, OH 43403, according to instructions elsewhere in this journal (see the table of contents). APA and the editors assume no responsibility for statements and opinions advanced by contributors to *Journal of Applied Psychology*.

Change of Address: Send change of address notice and a recent mailing label to the attention of the Subscription Section, APA, 30 days prior to the actual change of address. APA will not replace undelivered copies resulting from address changes; journals will be forwarded only if subscribers notify the local post office in writing that they will guarantee second-class forwarding postage.

Reprints: Authors may order reprints of their articles from the printer when they receive proofs.

Single Issues, Back Issues, and Back Volumes: For information regarding single issues, back issues, or back volumes write to Order Department, APA, 1200 Seventeenth Street, N.W., Washington, DC 20036.

Microform Editions: For information regarding microform editions write to either of the following: University Microfilms, Ann Arbor, MI 48106; or Princeton Microfilms, Princeton, NJ 08540.

Copyright and Permission: Authors must secure from APA and the author of reproduced material written permission to reproduce an article in full or text of more than 500 words. APA normally grants permission contingent upon like permission of the author, inclusion of the APA copyright notice on the first page of reproduced material, and payment of a fee of \$20 per page. Permission from APA and fees are waived for authors who wish to reproduce a single table or figure provided the author's permission is obtained and full credit is given to APA as copyright holder and to the author through a complete citation. Permission and fees are waived for authors who wish to reproduce their own material for personal use; fees only are waived for authors who wish to use more than a single table or figure of their own material commercially. Permission and fees are waived for the photocopying of isolated articles for nonprofit classroom or library reserve use by instructors and educational institutions. Access services may use abstracts without the permission of APA or the author. Libraries are permitted to photocopy beyond the limits of U.S. copyright law; (a) post-1977 articles, provided the per-copy fee in the code for this journal (0021-9010/87/\$00.75) is paid through the Copyright Clearance Center, 27 Congress Street, Salem, MA 01970; (b) pre-1978 articles, provided the per-copy fee stated in the Publishers' Fee List is paid through the Copyright Clearance Center, 27 Congress Street, Salem, MA 01970. Address requests for reprint permission to the Permissions Office, APA, 1200 Seventeenth Street N.W., Washington, DC 20036.

APA Journal Staff: Susan Knapp, *Executive Editor*; Leslie A. Cameron, *Director, Journals Program*; W. Ralph Eubanks, *Manager, Journal Production*; Lois Czapiewski and Theodore J. Baroody, *Production Editors*; Hugh Roberts, *Editorial Intern*; Jodi Ashcraft, *Advertising Sales Manager*.

The *Journal of Applied Psychology* (ISSN 0021-9010) is published quarterly (beginning in February) in one volume per year by the American Psychological Association, Inc., 1400 North Uhle Street, Arlington, VA 22201. Subscriptions are available on a calendar year basis only (January through December). The 1987 rates follow: *Non-member Individual*: \$60 Domestic, \$63 Foreign, \$70 Air Mail. *Institutional*: \$120 Domestic, \$127 Foreign, \$134 Air Mail. *APA Member*: \$30. Printed in the U.S.A. Second-class postage paid at Arlington, VA, and at additional mailing offices. POSTMASTER: Send address changes to the *Journal of Applied Psychology*, 1400 North Uhle Street, Arlington, VA 22201.

- 515 Effects of Exercise, Relaxation, and Management Skills Training on Physiological Stress Indicators: A Field Experiment
Nealia S. Bruning and David R. Frew
- 522 Impatience Versus Achievement Strivings in the Type A Pattern: Differential Effects on Students' Health and Academic Achievement
Janet T. Spence, Robert L. Helmreich, and Robert S. Pred
- 529 Occupational Complexity, Control, and Personal Income: Their Relation to Psychological Well-Being in Men and Women
Pamela K. Adelman
- 538 Understanding, Prediction, and Control as Moderators of the Relationships Between Perceived Stress, Satisfaction, and Psychological Well-Being
Lois E. Tetrick and James M. LaRocco
- 544 Pay, Equity, Job Gratifications, and Comparisons in Pay Satisfaction
Leonard Berkowitz, Colin Fraser, F. Peter Treasure, and Susan Cochran
- 552 Employee Age as a Moderator of the Relation Between Perceived Work Alternatives and Job Satisfaction
Samuel B. Pond III and Paul D. Geyer
- 558 Comparative Effects of Personal and Situational Influences on Job Outcomes of New Professionals
Stephen M. Colarelli, Roger A. Dean, and Constantine Konstans
- 567 Effect of Rater Training on Rater Accuracy: Levels-of-Processing Theory and Social Facilitation Theory Perspectives
Timothy R. Athey and Robert M. McIntyre
- 573 Behavioral Anchors as a Source of Bias in Rating
Kevin R. Murphy and Joseph I. Konstans
- 578 Measuring Occupational Difficulty: A Construct Validation Against Training Criteria
Michael D. Mumford, Joseph L. Weeks, Francis D. Harding, and Edwin A. Fleishman
- 588 Is Cost Accounting the Answer? Comparison of Two Behaviorally Based Methods for Estimating the Standard Deviation of Job Performance in Dollars With a Cost-Accounting-Based Approach
Olen L. Greer and Wayne F. Cascio
- 596 A Model of Hiring Decisions in Real Employment Interviews
Susan M. Raza and Bruce N. Carpenter
- 604 The Restriction of Range Problem and Nonignorable Selection Processes
Alan L. Gross and Mary Lou McGanney
- 611 Detecting Infrequent Deception
Kevin R. Murphy
- 615 Prosocial Behavior, Noncompliant Behavior, and Work Performance Among Commission Salespeople
Sheila M. Puffer

- 622 Business Climate Attitudes and Company Relocation Decisions
Neal Schmitt, Sandra E. Gleason, Bruce Pigozzi, and Philip M. Marcus
- 629 Improving the Reliability of Eyewitness Identification: Putting Context Into Context
Brian L. Cutler, Steven D. Penrod, and Todd K. Martens
- 638 Two (or More?) Dimensions of Organizational Commitment: Reexamination of the Affective and Continuance Commitment Scales
Gail W. McGee and Robert C. Ford
- 642 Some Time Dimensions of Work: Measurement of an Underlying Aspect of Organization Culture
Jacquelyn B. Schriber and Barbara A. Gutek
- 651 Improving Group Performance by Training in Individual Problem Solving
Preston C. Bottger and Philip W. Yetton
- 658 Comparative Analysis of Goal-Setting Strategies Across Cultures
Miriam Erez and P. Christopher Earley
- 666 Effect of Values on Perception and Decision Making: A Study of Alternative Work Values Measures
Elizabeth C. Ravlin and Bruce M. Meglino
- 674 Organizational Determinants of Leader Behavior and Authority
Tove H. Hammer and Jay M. Turk
- 683 Psychological Functioning Following an Acute Disaster
Julian Barling, Stephen D. Bluen, and Rolene Fain
- 691 Correlation of Eyewitness Accuracy and Confidence: Optimality Hypothesis Revisited
Robert K. Bothwell, Kenneth A. Deffenbacher, and John C. Brigham

Short Note

- 696 Predicting Supervisory Ratings Versus Promotional Progress in Test Validation Studies
Herbert H. Meyer

Other

- 698 Acknowledgment
- 528 Call for Nominations for *Journal of Experimental Psychology: General*
- 603 Call for Nominations for *Journal of Abnormal Psychology*
- 566 Delworth Appointed Editor of *Professional Psychology: Research and Practice*, 1989–1994
- 641 Instructions to Authors
- 557 Kintsch Appointed Editor of *Psychological Review*, 1989–1994
- 551 Schmitt Appointed Editor, 1989–1994

Effects of Exercise, Relaxation, and Management Skills Training on Physiological Stress Indicators: A Field Experiment

Nealia S. Bruning
Kent State University

David R. Frew
Gannon University

The physiological effects of three stress intervention strategies (management skills training, exercise, and meditation) were examined in a longitudinal field experiment. Sixty-two subjects were randomly assigned to four groups (three experimental groups and a control group). Pulse rate, diastolic blood pressure, systolic blood pressure, and galvanic skin response were used as physiological stress indicators. Analysis of covariance and multiple comparison tests indicated that each of the strategies led to decreases in pulse rate and systolic blood pressure. Dual combination strategies also showed significant decreases in pulse rate. However, no reliable results were found for combination strategies when examined for order effects.

The literature on job-related stress has grown geometrically during the past decade. Although the number of articles dealing with causes and effects of stress has increased, there has been relatively little work done in the area of stress intervention. To a large degree, this may be due to the lack of definitional clarity regarding stress itself. Much of the literature—including Abdel-Halim (1981), Beehr, Walsh, and Taber (1976), Burke and Belcourt (1974), Miles (1976), Parasuraman and Alutto (1981), Quick (1979), and Tosi (1971)—uses role clarity and other traditional job stressors as measures of stress. Rather than conceptualizing stress as a physiological or attitudinal response to stressful incidents at work, measures within these studies were simply descriptions of work circumstances. This shortcoming in the literature has been cited by Matteson and Ivancevich (1979) and has led to both theoretical and methodological confusion.

This research proceeds from the basis that work-related experiences create measurable stressors such as role ambiguity, excessive work demands, and interpersonal problems. These stressors may, in turn, cause measurable stress reactions in individual employees. In our view, the logical approach to examination of stress is the measurement of experienced stress by the employee.

This focus, however, leads to a second fundamental problem with the organizational literature on stress. Most of the existing literature that has attempted to measure experienced stress has focused on self-reported perceptions instead of on hard measures. Anderson (1976), Anderson, Hellriegel, and Slocum (1977) and Parkington and Schneider (1970) are examples of this methodology. The literature in physiology and medicine (e.g., Institute of Medicine, 1981) suggests that there may be substantial differences in coping abilities between individuals

who experience stressors. Persons who are less able to cope with stress would be expected to manifest this lower adaptivity in physiological measures of stress. The primary purpose of the research reported here was to investigate the impact of various intervention strategies on physiological indicators of stress.

According to Selye (1956), indicators of stress are measures that are indicative of overstimulation of the adrenal and autonomic systems in the body. Included in his list of indicators were adrenal outputs (e.g., epinephrine and norepinephrine) and autonomic outputs such as blood pressure (both systolic and diastolic), respiratory rate, pulse rate, galvanic skin response, focal activity, electrical brain activity, and glucose levels. Although both adrenal and autonomic responses are important, autonomic responses are generally easier to measure. Several authors have supported the measurement of autonomic responses to stress (e.g., Aro, 1984; Burke, 1980; Fowler, 1970; Gifford, 1975; Hogan & Hogan, 1982; Karasek, Theorell, Schwartz, Pieper, & Alfredsson, 1982; Kelleher, 1974; Wilkins, 1982). Research supporting autonomic system measurement is important because these responses have been the focus of most organizational studies that include physiological indicators (e.g., Gardner, 1982; Ivancevich, Matteson, & Preston, 1982; McGrath, 1976).

Intervention Strategies

A current controversy in the literature is related to stress intervention. Newman and Beehr (1979) noted, in their review of the literature on strategies for handling job stress, that there is a paucity of research in the area of intervention. Another difficulty noted by Newman and Beehr (1979) is the scarcity of well-designed and scientifically controlled studies in the intervention literature. They suggested that studies by Frew (1974), Caplan (1976), and Feinberg, Stabler, and Coley (1974) represented an interesting array of strategies for dealing with stress, but that each study suffered from the lack of a control group and a design that did not permit subsequent comparison of in-

Correspondence concerning this article should be addressed to Nealia S. Bruning, Department of Administrative Sciences, Kent State University, Kent, Ohio 44242.

interventions. Two recent reviews (McLeroy, Green, Mullen, & Foshee, 1984; Murphy, 1984) have indicated that the methodologies used to evaluate stress management strategies have improved greatly since the Newman and Beehr review. Major deficiencies of current studies include the lack of comparison between types of training and evaluation of combination programs (Murphy, 1984).

In a study that seems to represent the state-of-the-art approach to work-related stress intervention, Ganster, Mayes, Sime, and Tharp (1982) responded to one of the criticisms made by Newman and Beehr (1979). They carried out a well-controlled field experiment, using a treatment that consisted of both cognitive and relaxation training. Because of this single treatment, however, they were unable to compare the effects of different treatment modalities. A significant strength of the Ganster et al. study was that it used physiological criteria for measuring the stress variable.

Although Ganster et al. (1982) reported significant reductions in measured employee stress, they were reluctant to recommend either cognitive or relaxation training as a stress intervention tool. They felt that such a strategy might transfer the basic responsibility of experienced work stress from the organization to the individual. As a result, organizational efforts to reduce stressful job demands might be reduced.

The current state of the literature calls for a field experiment using physiological measures of stress as suggested by Ganster et al. (1982), but containing a series of separate and comparable treatments as suggested by Murphy (1984) and Newman and Beehr (1979). An overview of the Newman and Beehr typology suggested the following three distinct kinds of treatment strategies (two of which were combined in the Ganster et al. treatment). (a) *Management skills or cognitive management training*: Frequently, these types of programs are based on Meichenbaum's [1975] work on behavioral coping methods. The training may focus on such activities as goal setting, positive imagery, time management, communication skills, conflict resolution and problem articulation skills. (b) *Relaxation/meditation/biofeedback methods*: These methods are classified together because each requires a learned relaxation response or tension release. (c) *Exercise*: Exercise programs generally involve increased physical conditioning and aerobic activities. McLeroy et al. (1984) identified many more specific types of training, but the effectiveness of these three classes of programs is based on distinctively different underlying mechanisms. All fall within the McLeroy et al. definition of stress management techniques that do not focus on changing work site or job characteristics.

One hypothesis and three exploratory questions were formulated from a review of the literature discussed above. We hypothesized that all of the intervention strategies would result in improvements in physiological indicators of stress (i.e., reductions in pulse rate, diastolic and systolic blood pressure, and galvanic skin response). As a first exploratory issue we wanted to examine (given significant physiological improvements) the relative effectiveness of the three intervention strategies. A second exploratory issue was related to combination strategies: Do certain combinations of strategies lead to greater improvements

than other combinations? The third exploratory question examines the ordering effects of programs: Are there differences between combination programs based on their order of presentation? A longitudinal field experiment was used to test the hypothesis and provide information for the exploratory questions.

Method

Subjects

The study site was a hospital-equipment facility in northwestern Pennsylvania employing approximately 1,200 people within a three-location complex. The study subjects were selected from nonunion employees, a group of approximately 350 supervisors, managers, engineers, and technical and other support personnel. We approached corporate management and asked for permission to conduct the study/training sessions. There was not a clearly identified problem of experienced work stress, nor was the company actively seeking assistance in this area.

Participants were originally solicited at a lunchtime informational seminar on stress at work. Because the three plant locations were all close to one another, the information session and the subsequent training were all conducted in a central location. At the information meeting the proposed stress participants received a description of the requirements of the study. Of the 147 persons who attended the informational meeting, 86 volunteered to participate in the study. They were told that they would be assigned to four smaller training (treatment) classes and were subsequently (randomly) assigned to four separate groups. Of the original volunteers, 65 completed the 6-month study. Three of the volunteers had to be dropped from the study because of changes in medication: Two began medication programs after the initial measurements indicated dangerously high blood-pressure levels, and the third subject reduced his medication during the study. The remaining 21 persons who withdrew from the study did so for personal or job-related reasons such as transfer or turnover.

The subjects' ages ranged from 23 to 60 years ($M = 40.5$ years, $SD = 9.6$). Most of the subjects ($n = 55$) were married. The number of children ranged from 0 to 6; almost one third of the sample did not have children. Seventy-three percent had a college degree. About 80% of the subjects were men, and half of the subjects did not have subordinates. Average job tenure was 4 years, and average organizational tenure was 11 years.

Procedure

Each of the subjects volunteered to participate in an experiment on stress. They were randomly assigned to four groups (16 subjects in two groups, and 15 subjects in the remaining two groups). The four groups were as follows: (a) management skills (MSK), (b) meditation (MED), (c) exercise (EXC), and (d) control (CON). In the MSK group, the participants were instructed to explore both work and personal values and then to set both strategic and tactical goals. They were taught to pinpoint goals, to seek the collaboration of fellow workers regarding these objectives, and to identify roadblocks. After some instruction in the management of time and goal prioritization, the subjects were given communication training, including listening, skill development, and empathy. Much of this training program was based on Meichenbaum's (1975) work in behavioral coping methods.

The MED training was based on the progressive-relaxation and focused-meditation methods developed by Carrington (1978). Participants were instructed to sit quietly and meditate for 15 to 20 min either one or two times per day. They were instructed to begin each meditation

period with a few moments of calming and relaxing the body and then to spend the remaining time working with a mantra (meditation sound), which was selected by the participant from a list of such sounds taken from the work of Carrington.

Using the work of Cooper (1977), the EXC group was instructed to spend 30 min every other day (three ½-hr periods per week) participating in aerobic (not anaerobic) activities of their choice, including walking, running, swimming, bicycling, and stationary bicycling. Conservative guidelines were used in the instruction. The subjects were advised to raise their pulse rates during exercise by at least 15% of the normal resting rate, but by no more than 75% of the computed value of 220 minus the subject's age. The subjects were further advised to alternate activities if they wished (e.g., walking and swimming).

The control group was given general information pertaining to stress. In all three training groups, the participants were taught a strategy that they were expected to practice. The respondents kept individual logs of their practice rates that were handed into the experimenters at the end of the experiment. These logs served as a reminder to keep practicing the techniques. Physiological measures were taken before beginning the experiment. These measures were then taken at approximately 2-week intervals for 6 months. Only three of these measurements were incorporated into this study: the 1st (A), 6th (B), and 10th (C) measurements. After the first assessment of physiological measures, all the groups, with the exception of the control group, participated in 8–10 hr of training during a regular work week (during lunch and working hours). All training groups were taught the methods of exercise, relaxation, and management skills. The subjects did not practice the methods extensively in the training session. For example, the EXC group was taught how to monitor their heart rate and the aerobic value of various exercises, but they did not exercise during the training sessions. Thirteen weeks later, the Time B measurement of physiological indicators was taken on all four groups. At this point the three original treatment groups were split and assigned a second treatment (receiving an additional 8–10 hr of training). The CON group received all three types of training. (Figure 1 illustrates this second assignment.) Approximately 10 weeks later the C measurement of the physiological indicators was taken. The same nurses took all of the physiological measurements. A single trainer (one of the investigators) conducted all of the training sessions. Table 1 summarizes the study design.

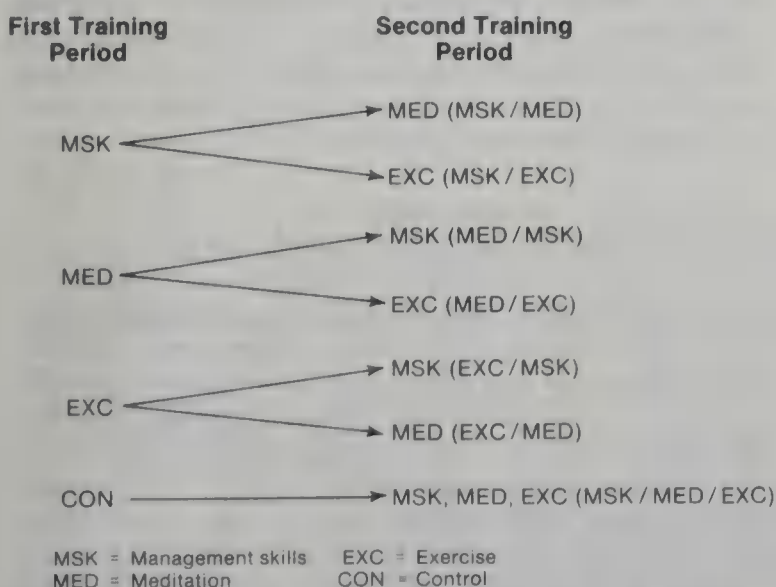


Figure 1. Training assignments for the first and second training periods.

Table 1
Study Design

Stages of research	Group			
	1	2	3	4
Assignment to groups	Random	Random	Random	Random
Measurement episode (October)	A	A	A	A
8-week treatment	MSK	MED	EXC	CON
Measurement episode (January)	B	B	B	B
6-week treatment ^a	MED, EXC	MSK, EXC	MSK, MED	ALL
Measurement episode (April)	C	C	C	C

Note. A = Time 1; B = Time 2; C = Time 3. MSK = cognitive skills treatment; MED = relaxation treatment; EXC = exercise treatment; and CON = Control group.

^a During the 6-week treatment, original groups were split into two subgroups, and subjects were each given a second treatment (MSK + MED, MSK + EXC, MED + MSK, etc.). The control group (CON) was given all three treatments.

Instruments

Four separate physiological measures that have been associated with stress and strain (e.g., Burke, 1980; Fowler, 1970; Gifford, 1975; Kelleher, 1974; Selye, 1956) were taken. These indicators were pulse rate (PULSE), diastolic (DIAS) and systolic (SYS) blood pressure, and galvanic skin response (GSR). The blood pressure and pulse readings were supervised by a registered nurse, and the blood-pressure readings were taken from a recently calibrated set of two standard Aneroid Sphygmomanometers. The same nurses, who were blind to the treatments, performed all of the measurements. Subjects were tested between 10 and 11 a.m. in the same location. Each participant was instructed to sit quietly for 5 min prior to measurement. The GSR readings were taken by one of the primary investigators using an Edmund Scientific Model 508 monitor. At the beginning of the study, these four physiological indicators were measured 1 week apart to get a measure of reliability. The test-retest reliabilities were as follows: PULSE, .62; DIAS, .82; SYS, .79; and GSR, .92. The correlations between these physiological measures were generally below .18. The only exception was between DIAS and SYS ($r = .53, p < .05$).

Data Analysis

Two sets of one-way analyses of covariance (ANCOVAs) were used to determine changes in the four physiological measures over time. The first set of analyses was an examination of the effects (shown by the physiological indicators) of the three treatment groups after 3 months (Time B). The Time-A physiological measurement was included as a covariate in the ANCOVA to control for initial individual differences and for systematic differences owing to certain personal characteristics (e.g., sex). According to our first hypothesis, the treatment groups should result in improvements in the physiological measurements. We examined the differences between the treatment groups as an exploratory issue. The exploratory analysis was conducted by using a Newman-Keuls post

Table 2
Unadjusted Means and Standard Deviations for the Treatment Conditions and Control Group at Times 1 and 2

Physiological variable	MSK		MED		EXC		CON	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
PULSE A	74.75	9.82	74.40	11.77	71.87	12.33	75.62	13.95
DIAS A	93.81	11.29	85.13	14.81	83.33	6.18	80.56	11.35
SYS A	139.38	12.03	135.47	10.12	129.13	9.20	135.94	13.44
GSR A	37.19	19.74	35.33	14.66	26.80	12.49	26.19	13.12
PULSE B	64.50	8.47	69.87	10.62	68.73	9.37	79.00	10.30
DIAS B	80.56	8.75	78.13	10.62	74.13	4.56	78.62	8.51
SYS B	126.56	7.84	126.00	7.52	121.33	5.38	132.25	10.25
GSR B	20.94	13.45	24.33	14.57	19.67	12.80	23.00	12.93

Note. MSK = management skills; MED = meditation; EXC = exercise; CON = control. PULSE = pulse rate; DIAS = diastolic blood pressure; SYS = systolic blood pressure; GSR = galvanic skin response. A = Time 1; B = Time 2; C = Time 3. *N*s for the various treatments are as follows: MSK = 16; MED = 15; EXC = 15; CON = 16. A = Time 1; B = Time 2.

hoc multiple comparison analysis on the measures that had a significant ($p < .05$) *F* value (Glass & Hopkins, 1984).

The second set of analyses used the Time-C measures as the dependent variables and the Time-A physiological measurements as the covariates. In these ANCOVAs, four groups were analyzed: the two MSK/MED groups combined (i.e., both the MSK → MED treatment and the MED → MSK treatment), the two MSK/EXC groups combined, the two MED/EXC groups combined, and the CON/MSK/MED/EXC group. This set of analyses, for which no research hypothesis was proposed, was conducted to test the combination effects of the programs. These ANCOVAs were designed to answer the second exploratory question: Do certain combination strategies lead to greater improvements in physiological measures than other combination strategies?

The third set of analyses was structured to address the third exploratory question: Were there ordering effects in the programs? The seven treatment groups described in the Procedure section were the independent variables, the Time-A physiological measures were the covariates, and the Time-C physiological measures served as the dependent variables. As in the first set of analyses, the Newman-Keuls procedure was used in the last two sets of analyses to test the differential effects (in the physiological measures) of the different combination treatment strategies.

Results

The unadjusted means and standard deviations for the treatment conditions and the control groups may be found in Tables 2 and 3. Upon examining these figures, It was apparent that some of the Time-A means of a given variable across groups may have been significantly different from one another. Therefore, a one-way analysis of variance (ANOVA) was conducted for each of the four physiological measures. The only ANOVA that was statistically significant was for DIAS, $F(3, 58) = 4.06$, $p < .05$.

The results pertaining to our first hypothesis can be found in Table 4. Examination of the top section of the table indicates that treatment effects were significant ($p < .05$) for two of the physiological measures, PULSE, $F(3, 57) = 9.59$, $p < .01$, and SYS, $F(3, 57) = 5.15$, $p < .01$. Therefore, we found support for the first hypothesis for two of the four stress indicators.

The middle section of Table 4 summarizes the results due to the combination training methods. There was a significant decline for PULSE, $F(3, 57) = 3.69$, $p < .05$, but not for the other dependent variables. The bottom section of Table 4 summa-

Table 3
Unadjusted Means and Standard Deviations for the Combined Treatment Conditions and Control Group at Time 3

Physiological variable	MSK → MED		MSK → EXC		MED → MSK		MED → EXC		EXC → MSK		EXC → MED		CON → MSK MED EXC	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
PULSE C	65.88	4.09	64.25	10.98	66.88	10.95	70.29	5.47	66.25	7.59	62.29	10.55	72.62	7.68
DIAS C	80.75	6.58	76.25	6.71	7.00	8.28	77.14	7.99	70.38	3.85	71.71	6.78	74.38	8.14
SYS C	125.50	9.24	126.25	9.77	125.25	7.40	123.71	11.22	123.75	4.20	116.86	7.65	125.19	8.95
GSR C	24.25	17.21	14.62	8.28	18.25	16.41	22.57	9.96	16.50	10.50	14.57	3.78	17.62	14.24

Note. MSK = management skills; MED = meditation; EXC = exercise; CON = control. PULSE = pulse rate; DIAS = diastolic blood pressure; SYS = systolic blood pressure; GSR = galvanic skin response. C = Time 3. *N*s for the various treatments are as follows: MSK/MED, MSK/EXC, MED/MSK, EXC/MSK = 8 in each group; MED/EXC, EXC, MED = 7 in each group; CON/MSK/MED/EXC = 16.

Table 4
*Analysis of Covariance with Physiological Measures
 as Dependent Variables and Time A
 Physiological Measure as Covariate*

Dependent variable	Covariate	Covariate F	df	Treatment F	df
Effects owing to individual training methods					
PULSE B	PULSE A	44.36**	1,57	9.59**	3
DIAS B	DIAS A	45.60**	1,57	2.38	3
SYS B	SYS A	58.41**	1,57	5.15**	3
GSR B	GSR A	40.52**	1,57	1.70	3
Effects owing to combination training methods					
PULSE C	PULSE A	56.80**	1,57	3.69*	3
DIAS C	DIAS A	58.38**	1,57	2.17	3
SYS C	SYS A	37.51**	1,57	0.96	3
GSR C	GSR A	52.92**	1,57	0.43	3
Effects owing to ordering of the combination training methods					
PULSE C	PULSE A	59.74**	1,54	2.93*	6
DIAS C	DIAS A	58.92**	1,54	1.68	6
SYS C	SYS A	36.05**	1,54	0.59	6
GSR C	GSR A	50.82**	1,54	0.92	6

Note. PULSE = pulse rate; DIAS = diastolic blood pressure; SYS = systolic blood pressure; GSR = galvanic skin response; A = Time 1; B = Time 2; C = Time 3.

* $p < .05$. ** $p < .01$.

rizizes the ordering effects of the combination treatments. One of the overall F s was statistically significant, PULSE, $F(6, 54) = 2.93$, $p < .05$. Because we did not intend to test any specific hypotheses in these second ANCOVA results, no conclusions were derived at this point in the analyses.

The Newman-Keuls analyses (Table 5) provided information pertinent to the exploratory questions, that is, more specific information about the differences between the treatment conditions. The comparison between the treatment groups and the control group for pulse rate indicates that all three interventions resulted in significant differences. Significant reductions in SYS were found for all treatment conditions, but no differences were found between treatment conditions. These results were also supported using the more conservative Tukey test (see the note to Table 5).

The results for the combination training strategies and the ordering effects are also summarized in Table 5. PULSE was the only variable on which these analyses were conducted. The Q values indicate that all the combination strategies led to reductions in PULSE, but no combination strategy was better than another. However, two of these differences were no longer significant when tested using the Tukey procedure.

The results for the ordering effects provided evidence that three of the ordered programs led to decreases in PULSE. One of these differences was no longer significant when the Tukey procedure was used. The adjusted mean for the MED/EXC was also significantly higher than the means for the EXC/MED group and the MSK/EXC group.

Discussion

In this study a longitudinal field experiment was designed to examine the relative effectiveness of three intervention strategies: management skills training, meditation, and exercise. The hypothesis that stress intervention strategies would improve physiological indicators was supported for two of four indicators (pulse rate and systolic blood pressure).

The combination strategy analyses indicated that all combination groups led to significant decreases in pulse rate, but none of the combination strategies appeared to be superior to others. Finally, the ordering appeared to make some difference in comparative decreases in pulse rates, but these differences may not be replicated in future studies.

Given the number of post hoc comparisons conducted, one might criticize these conclusions on the basis of experiment-wise error. The more conservative Tukey test that was applied to the data would change the conclusion for four of the mean differences. All of the individual training methods led to decreases in pulse rate and systolic blood pressure even with the more stringent criterion.

Several observations are apparent from this study. First, pulse rate seemed to be the most sensitive physiological measure. Therefore, a pertinent question for this kind of research might focus on the informational value of pulse rate compared with the other measures used. Do reductions in pulse rate possibly lead to later improvements in other vital signs? If this is the case, the ease with which pulse rate can be measured could lead to a great deal of related and valuable organizational research.

One criticism of this study is the loss of the control group after the second time period. Pressure from the participants for training prohibited the investigators from maintaining a control group throughout the experiment. However, because decreases in the pulse rates were still evident at Time 3, the loss of the control group may have led to more conservative conclusions than would have been apparent if the control group had been maintained. Differences between combination strategies and ordering effects may not have been significant in these analyses even if the control group had been retained.

Another criticism of this study concerns the value of physiological indicators of stress. Although many earlier researchers have supported their use, other researchers might have led us to interpret these results more conservatively. For example, in a recent review, evidence was presented that changes in physiological measures may be evidence of alert excitement rather than of longer term pathological changes (Light, 1981). However, the results reported here indicate consistent reductions in pulse rates owing to training, and pulse rate is an indicator of cardiovascular functioning (Wilkins, 1982).

What are the organizational implications of these findings? Edwards and Gettman (1980) raised questions about the monies spent on employee physical fitness programs without clear indicators of their cost effectiveness. From the present results, it would appear that each of several types of intervention strategies may have at least some beneficial effects. Cost considerations might suggest that trainers should examine various treatment options because less expensive options (e.g., management skills training ver-

Table 5
Newman-Keuls Analyses for Significant Analysis of Covariance Results

Dependent variable/ treatment condition	<i>M</i>		Group <i>Q</i> values					
	Group ^a	Order ^b						
Effects owing to individual training methods								
PULSE B			CON	EXC	MED			
MSK	64.21	CON	—	—	—			
MED	69.75	EXC	4.34**	—	—			
EXC	69.96	MED	4.44**	.10	—			
CON	78.15	MSK	7.38**	3.03	2.93†			
SYS B								
MSK	125.53	CON	—	—	—			
MED	125.81	MED	3.99**	—	—			
EXC	124.14	MSK	4.17**	.18	—			
CON	131.84	EXC	5.09**	1.10	.92			
Effects owing to combination training methods								
PULSE C			MSK/MED/EXC	MED/EXC	MSK/MED			
MSK/MED	66.50	MSK/MED/EXC	—	—	—			
MSK/EXC	65.13	MED/EXC	3.20†	—	—			
MED/EXC	67.08	MSK/MED	3.59†	.38	—			
CON/MSK/ MED/EXC	71.93	MSK/EXC	4.49*	1.29	.90			
Effects owing to ordering of combination training method								
PULSE C			MSK/MED/EXC	MED/EXC	EXC/MSK	MED/MSK	MSK/MED	EXC/MED
MSK/MED	66.40	MSK/MED/EXC	—	—	—	—	—	—
MSK/EXC	63.19	MED/EXC	1.10	—	—	—	—	—
MED/MSK	66.61	EXC/MSK	3.44†	2.34	—	—	—	—
MED/EXC	70.38	MED/MSK	3.78	2.68	.33	—	—	—
EXC/MSK	67.08	MSK/MED	3.93	2.83	.48	.01	—	—
EXC/MED	63.78	EXC/MED	5.79**	4.69*	2.34	2.01	1.86	—
MSK/MED/EXC	71.93	MSK/EXC	6.21**	5.10*	2.76	2.43	2.28	.42

Note. Analysis of covariance was considered significant at the .05 level. MSK = management skills; MED = meditation; EXC = exercise; CON = control.
^a Group M adjusted for covariate. ^b This column lists the treatments for each variable, starting with the treatment giving the lowest mean value of the variable. The treatment here does not refer to any characteristic of the treatment at the beginning of the same row.
* $p < .05$. ** $p < .01$. † Using the Tukey method of multiple comparisons, with a family error rate of $p < .05$, these Q values would not be significant.

sus elaborate gym facilities) might lead to similar results. The present results seem to indicate that physiological changes may be either improved through cognitive changes or through better physical conditioning, but there is some evidence to suggest that cognitive changes may be more beneficial (on the basis of the Time-B multiple comparisons for pulse rate).

In conclusion, these results provide partial support for the introduction of stress intervention programs. More evaluation is needed regarding the relative effectiveness of strategies and the cost effectiveness of various intervention programs. It is our view that stress intervention strategies may offer the individual effective methods to reduce the negative effects of stress and, if so, that organizations should make an effort to encourage employees to participate in these programs. This does not, as Ganster et al. (1982) warned, relieve the organization

of its responsibility to alleviate the root causes of work-related stress, but it may provide short- and long-term benefits to employees.

References

Abdel-Halim, A. (1981). Effects of role stress-job design-technology interaction on employee work satisfaction. *Academy of Management Journal*, 24, 260-273.

Anderson, C. (1976). Coping behaviors as intervening mechanisms in the inverted U stress performance relationship. *Journal of Applied Psychology*, 61, 30-34.

Anderson, C., Hellriegel, D., & Slocum, J. (1977). Managerial response to environmentally induced stress. *Academy of Management Journal*, 20, 260-272.

- Aro, S. (1984). Occupational stress, health-related behavior, and blood pressure: A 5-year follow-up. *Preventive Medicine, 13*, 333-348.
- Beehr, T., Walsh, J., & Taber, T. (1976). Relationship of stress to individually and organizationally valued stress: Higher order needs as a moderator. *Journal of Applied Psychology, 61*, 41-47.
- Burke, S. (1980). *Human anatomy and physiology for the health sciences*. New York: Wiley.
- Burke, R., & Belcourt, M. (1974). Managerial role stress and coping responses. *Journal of Business Administration, 5*, 55-68.
- Caplan, R. D. (1976, August). *Occupational differences in job demands and strain*. Paper presented at the 84th annual meeting of the American Psychological Association, Washington, DC.
- Carrington, P. (1978). *Freedom in meditation*. New York: Anchor Press/Doubleday.
- Cooper, K. H. (1977). *The aerobics way*. New York: M. Evans.
- Edwards, S., & Gettman, L. (1980). The effect of employee fitness on job performance. *Personnel Administrator, 25*(11), 41-44, 61.
- Feinberg, A. M., Stabler, B., & Coley, S. B. (1974). Electrically induced relaxation in systematic desensitization: A case note. *Psychological Reports, 35*, 75-78.
- Fowler, N. (1970). *Inspection and palpation of venous and arterial pulses*. New York: American Heart Association.
- Frew, D. (1974). Transcendental meditation and productivity. *Academy of Management Journal, 17*, 362-368.
- Ganster, D. C., Mayes, B. T., Sime, W. E., & Tharp, G. D. (1982). Managing organizational stress: A field experiment. *Journal of Applied Psychology, 67*, 533-542.
- Gardner, D. G. (1982). An empirical test of activation theory predictors about the effects of task design on major job strain indices. *Proceedings of the 13th annual Midwest American Institute for Decision Sciences*, 16-18.
- Gifford, R. (1975). Hypertension 1975. *Drug Therapy*, May/June, 5-9.
- Glass, G., & Hopkins, K. (1984). *Statistical methods in education and psychology*. Englewood Cliffs, NJ: Prentice Hall.
- Hogan, R., & Hogan, J. C. (1982). Subjective correlates of stress and human performance. In E. A. Alluisi & E. A. Fleishman (Eds.), *Human performance and productivity: Stress and performance effectiveness* (pp. 141-163). Hillsdale, NJ: Erlbaum.
- Institute of Medicine (1981). *Research on stress and human stress* (Report No. PB82-249095). Springfield, VA: National Technical Information Service, U.S. Department of Commerce.
- Ivancevich, J. M., Matteson, M. T., & Preston, C. (1982). Occupational stress, type A behavior and physical well being. *Academy of Management Journal, 25*, 373-391.
- Karasek, R. A., Theorell, T. G. T., Schwartz, J., Pieper, C., & Alfredsson, L. (1982). Job, psychological factors and coronary heart disease. *Advances in Cardiology, 29*, 62-67.
- Kelleher, M. (1974). Systolic pressure in assessing treatment effects. *The New England Journal of Medicine, 291*, 1192-1193.
- Light, K. C. (1981). Cardiovascular responses to effortful active coping: Implications for the role of stress in hypertension development. *Psychophysiology, 18*, 216-225.
- Matteson, M., & Ivancevich, J. (1979). Organizational stressors and heart disease. *Academy of Management Review, 4*, 347-357.
- McGrath, J. (1976). Stress and behavior in organizations. In M. Dunnette (Ed.), *Handbook of industrial and organizational psychology*. Chicago: Rand McNally.
- McLeroy, K. R., Green, L. W., Mullen, K. D., & Foshee, V. (1984). Assessing the effects of health promotion in worksites: A review of the stress program evaluations. *Health Education Quarterly, 11*, 379-401.
- Meichenbaum, D. H. (1975). A self instructional approach to stress management. In C. D. Spielberger & J. H. Sarason (Eds.), *Stress and anxiety* (Vol. 1, pp. 237-263). New York: Halstead Press.
- Miles, R. H. (1976). Role requirements as sources of organizational stress. *Journal of Applied Psychology, 61*, 172-179.
- Murphy, L. R. (1984). Occupational stress management: A review and appraisal. *Journal of Occupational Psychology, 57*, 1-15.
- Newman, J., & Beehr, T. (1979). Personal and organizational strategies for handling job stress: A review of research and opinion. *Personnel Psychology, 32*, 1-43.
- Parasuraman, S., & Alutto, J. A. (1981). An examination of the organizational antecedents of stressors at work. *Academy of Management Journal, 24*, 48-67.
- Parkington, J., & Schneider, B. (1979). Some correlates of experienced job stress. *Academy of Management Journal, 22*, 270-281.
- Quick, J. C. (1979). Dyadic goal setting and role stress: A field study. *Academy of Management Journal, 22*, 241-252.
- Selye, H. (1956). *The stress of life*. New York: McGraw-Hill.
- Tosi, H. (1971). Organization stress as a moderator of the relationship between influence and role response. *Academy of Management Journal, 14*, 7-20.
- Wilkins, W. L. (1982). Psychophysiological correlates of stress and human performance. In E. A. Alluisi & E. A. Fleishman (Eds.), *Human performance and productivity: Stress and performance effectiveness* (pp. 57-90). Hillsdale, NJ: Erlbaum.

Received July 1, 1985

Revision received March 26, 1987

Accepted May 4, 1987 ■

Impatience Versus Achievement Strivings in the Type A Pattern: Differential Effects on Students' Health and Academic Achievement

Janet T. Spence, Robert L. Helmreich, and Robert S. Pred
University of Texas at Austin

Psychometric analyses of college students' responses to the Jenkins Activity Survey, a self-report measure of the Type A behavior pattern, revealed the presence of two relatively independent factors. On the basis of these analyses, two scales, labeled *Achievement Strivings* (AS) and *Impatience-Irritability* (II), were developed. In two samples of male and female college students, scores on AS but not on II were found to be significantly correlated with grade point average. Responses to a health survey, on the other hand, indicated that frequency of physical complaints was significantly correlated with II but not with AS. These results suggest that there are two relatively independent factors in the Type A pattern that have differential effects on performance and health. Future research on the personality factors related to coronary heart disease and other disorders might more profitably focus on the syndrome reflected in the II scale than on the Type A pattern.

On the basis of their observations of patients with coronary heart disease (CHD), Friedman and Rosenman (1959, 1974) proposed that proneness to CHD is associated with a behavior pattern that they labeled *Type A*. In their formulation, people who exhibit the Type A pattern are characterized by ambitiousness, competitiveness, time urgency, impatience, and aggressiveness or hostility. Individuals who are relatively lacking in these characteristics are identified as *Type B*.

Several assessment devices have been developed to classify individuals as Type A or B, the most commonly used being the Structured Interview technique (SI; Rosenman et al., 1964; Rosenman, 1978) and an objectively scored self-report measure, the Jenkins Activity Survey for Health Predictions (JAS; Jenkins, Zyzanski, & Rosenman, 1971), which was designed to tap the same characteristics as the SI. A student form of the JAS in which items referring to job or job setting have been eliminated or modified (by substituting references to school work and the academic setting) has also been developed (Krantz, Glass, & Snyder, 1974). Studies using these several devices not only provide evidence for a link between the Type A pattern and CHD, but also suggest that this pattern may be associated with related disorders (e.g., Haynes, Feinleib, Levine, Scotch, & Kannel, 1978; Manuck, Morrison, Bellack, & Polefrone, 1985).

Many of the components of the Type A pattern involve achievement-related motives and behaviors commonly believed to contribute to successful academic and vocational performance. Although Friedman and Rosenman (1974) have suggested that time urgency and emphasis on quantity rather than quality of work in Type As may interfere with effective perfor-

mance, general discussions of the Type A concept typically imply that hard-driving, achievement-oriented people classified as Type A are more likely to succeed than the more relaxed, less ambitious people classified as Type B. Jenkins, Zyzanski, and Rosenman (1971) stated, for example, that "Individuals with the [Type A] pattern are usually conscientiously committed to their occupation, and whatever its level, often have achieved success in it" (p. 194).

The evidence, although sparse, supports the view that, as a group, Type As outperform Type Bs. Thus, Type A college students have been found to earn more academic honors (Glass, 1977) and higher grades (Waldron et al., 1980) than Type Bs. In another investigation, Matthews, Helmreich, Beane, and Lucker (1980) obtained JAS data from a subset of male academic psychologists studied by Helmreich, Spence, Beane, Lucker, and Matthews (1980). Matthews et al. (1980) reported positive correlations between the respondents' JAS scores and two measures of scholarly attainment: number of publications and number of citations by others to their work (i.e., those high in Type A behaviors tended to be more productive and more highly cited than their Type B peers). Greater productivity in Type A faculty members has also been found by Taylor, Locke, Lee, and Gist (1984). The uncomfortable conclusion suggested by these findings is that persons exhibiting the Type A pattern are likely to be more successful than those classified as Type B but that they simultaneously risk paying a heavy price for their attainments in terms of CHD or other health problems.

It seems highly unlikely that the same components of the Type A pattern are responsible for the positive association with indexes of vocational and academic excellence, on the one hand, and with the greater incidence of CHD and other health problems, on the other hand. Common sense suggests that the positive relation with accomplishments is brought about by the achievement strivings of Type As per se. The empirical literature provides indirect support for this contention. For example, in the Matthews et al. (1980) study of academic psychologists, measures of mastery and work-oriented achievement motives

This research was sponsored, in part, by Cooperative Agreement NCC2-286 between the National Aeronautics and Space Administration and the Ames Research Center, Moffet Field, California (Robert L. Helmreich).

Correspondence concerning this article should be addressed to Janet T. Spence, Department of Psychology, University of Texas at Austin, Austin, Texas 78712.

(Spence & Helmreich, 1978) were positively related both to JAS scores and to the productivity and citation measures.

As for the components responsible for the relations with CHD and other associated disorders, a number of investigators (e.g., Matthews, 1982; Rosenman, 1985; Spielberger et al., 1985; Williams et al., 1980) have recently pointed to what Spielberger et al. (1985) have labeled the *AHA! Syndrome*: anger, hostility, and aggression. The deleterious effects of this set of interrelated variables on health have been independently established (e.g., Diamond, 1982; Greer & Morris, 1975). Direct support for the contention that it is these elements in the Type A pattern that may be uniquely responsible for the association between Type A behavior and CHD comes from a study by Matthews, Glass, Rosenman, and Bortner (1977). These investigators reanalyzed SI data from the Western Collaborative Group Study, a large prospective project, which found that middle-aged men initially classified as Type As were more likely to develop CHD than those classified as Type Bs (Rosenman et al., 1964). Matthews et al. (1977) reported that the SI items most sharply distinguishing CHD cases from non-CHD cases were those related to irritability, anger, hostility, and several behaviors that could be motivated by these variables (e.g., vigorous answers and explosive voice modulation during the interview). Similar findings have recently been reported by Weinstein, Davison, DeQuattro, and Allen (1986).

The data thus suggest that different aspects of the Type A pattern may be responsible for the correlations of this pattern with academic and vocational performance and with measures of CHD and other medical conditions. If the evidence continues to indicate that only some components of the Type A pattern influence health and it can be demonstrated that a different set of components influences performance, the question that must be asked is whether the Type A construct is a useful one. That is, can it be assumed that the various components assigned to the Type A pattern show a strong tendency to coexist? Put another way, can persons who are hard-driving, hard working, and ambitious usually be characterized as irritable, impatient, and hostile as well?

Unless the various facets of the Type A pattern can be demonstrated to have similar effects on various types of outcome measures, or these facets are shown to be highly correlated, it would seem to be more profitable to abandon the Type A concept. Those interested in the personality and behavioral factors related to CHD and other related disorders might better turn their attention away from the Type A pattern and consider directly the *AHA! Syndrome* per se. Similarly, those interested in the personal factors contributing to successful vocational and academic performance should look to sets of characteristics directly related to achievement strivings rather than to the Type A pattern. Additional evidence is required, however, before the usefulness of the Type A concept can be seriously disputed.

In an effort to address these issues, we conducted psychometric analyses of the JAS data obtained from male and female college students on the student form of the JAS (Pred, Helmreich, & Spence, in press). Factor analyses with oblique rotations, based on unit-weighting of the individual items, yielded similar two-factor solutions in both sexes. Items with heavy loadings on the first factor appear to be related to achievement-related strivings (e.g., hard working, active, takes work seri-

ously). Those with heavy loadings on the second factor appear to tap impatience, irritability, and anger. In both sexes, the correlation between the two factors was low (about .15), thus suggesting that there is considerable independence between them. On the basis of the results of these analyses, two unit-weighted factor scales were constructed and labeled *Achievement Strivings* (AS) and *Impatience-Irritability* (II). Confirmatory factor analyses performed on data from the Matthews et al. (1980) psychologist sample and from another sample of students given the AS and II scales verified these item assignments (Pred et al., in press).

The purpose of the study reported here was to determine whether these two facets of the Type A pattern have differential effects. Students' scores on the AS and II scales were related to their grade point average (GPA) with the expectation that the relation with AS, but not with II, would be significant. In addition, the subjects were asked to respond to a health survey in which they were queried about sleep disturbances, respiratory disorders, headaches, and digestive upsets. In this instance, it was anticipated that II would be a better predictor of ill-health than AS.

Method

Subjects

Data were obtained from two samples of college students enrolled in introductory psychology courses at the University of Texas at Austin during the fall and spring semesters of the 1985-1986 academic year. The first sample consisted of 362 men and 351 women. These students were given the 44-item student version of the JAS developed by Krantz et al. (1974). The second sample was composed of 256 men and 225 women. They were given (in revised form) only the items assigned to the new AS and II scales. In the combined samples, 67% were classified as freshmen, 19% as sophomores, 9% as juniors, and 5% as seniors. All subjects participated as part of a course requirement.

Measures

Jenkins Activity Survey (JAS). Items on the 44-item JAS given to the first sample are accompanied by rating scales with 2 to 5 points. The items have been broken down by Jenkins et al. (1971) into three subscales with overlapping content. The first is the 21-item A-B scale, whose items were selected and assigned different weights by means of discriminant function analyses designed to maximize the concordance between the Type A classification produced by the Structured Interview and by the JAS. The other two are a priori factor scales: Factor H (Hard-driving competitiveness) and Factor S (Speed and impatience). The H and S scales consist of 17 and 21 items, respectively. Item overlap is particularly marked between the A-B and each of the two factor scales, but the latter also have four items in common.

The Factor H and S scales and the A-B scale are scored by optimal weights. Other investigators (e.g., Glass, 1977) however, have used a unit-weighting system. Furthermore, Matthews et al. (1980) reported that unit weighting of their JAS data yielded scores whose correlation with scores produced by the discriminant-function weighting was .90. For this reason, we used a unit-weighting system in this investigation. However, the scores assigned individual items were adjusted in an attempt to give the items with different numbers of alternatives more equal weight. For all items, a score of 5 was given to the extreme Type A response. For 2-point scales, the remaining (i.e., non-Type A) response was scored 2.5, whereas for 3-point scales, the remaining scores

Table 1
Items on the Revised Achievement Strivings and Impatience-Irritability Scales

Achievement Strivings
1. How much does college “stir you into action?” (<i>much less to much more than others</i>)
2. Nowadays, do you consider yourself to be: (<i>very hard-driving to very relaxed and easy going</i>)
3. How would your best friends or others who know you well rate your general level of activity? (<i>too slow to very active; should slow down</i>)
4. How seriously do you take your work? (<i>much more to much less than most</i>)
5. How often do you set deadlines or quotas for yourself in courses or other activities? (<i>very often to almost never</i>)
6. Compared with other students, the amount of effort I put forth is: (<i>much more to much less</i>)
7. Compared with other students, I approach life in general: (<i>much more to much less seriously</i>)
Impatience-Irritability
1. When a person is talking and takes too long to come to the point, how often do you feel like hurrying the person along? (<i>very frequently to almost never</i>)
2. Typically, how easily do you get irritated? (<i>extremely easily to not at all easily</i>)
3. Do you tend to do most things in a hurry? (<i>definitely true to not at all true</i>)
4. How is your “temper” these days? (<i>very hard to control to I seldom get angry</i>)
5. When you have to wait in line such as at a restaurant, the movies, or the post office, how do you usually feel? (<i>accept calmly to feel very impatient and refuse to stay long</i>)

Note. The labels for the end points of the 5-point rating scale accompanying each item appear in abbreviated form in parentheses.

were 3.33 and 1.67 and for 4-point scales they were 3.75, 2.5, and 1.25. For 5-point scales, scores ranged from 5 to 1.

As reported in Pred et al. (in press), the JAS data were subjected to factor analyses, separately for each sex, using a principal-axis solution, and to an oblique rotation, using an oblimin solution (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975). On the basis of an eigenvalue-one criterion, a two-factor solution was subjected to oblique rotation. For men, using a .35 criterion, 8 items loaded only on the first factor, 5 items only on the second factor, 3 items on both factors, and 28 items on neither factor. For women, the corresponding numbers were 10, 5, 1, and 28. For both sexes, items loading on the first factor reflected achievement-related behaviors and attitudes (e.g., hard-driving, puts forth much effort, takes work seriously). As mentioned above, this factor was labeled *Achievement Strivings* (AS). Items loading on the second factor described impatience, irritability, and anger. This factor was labeled *Impatience-Irritability* (II). The factor correlations were .16 and .14 for men and women, respectively.

Seven of the items reaching the .35 loading criterion in both sexes on the first factor were assigned to an AS scale. The five items (which were the same for both sexes) reaching this criterion on the second factor were assigned to an II scale. Scores on these two new scales were then found for each subject. The Cronbach alphas on the AS scale were .69 and .72 for men and women, respectively. Corresponding alphas for the II scale were .65 and .64.

In the second sample, subjects were given a revised version of the AS and II scales. The major revision was an expansion of the rating scales accompanying each item to a 5-point scale. A number of items were also slightly reworded, primarily to accomodate the new rating scales. The data were subjected to confirmatory factor analyses, separately for each sex (Pred et al., in press). For both sexes, the analysis replicated the two-factor structure obtained in the first sample: All items originally assigned to the AS scale loaded .30 or greater on this factor, and all items originally assigned to the II scale loaded .30 or greater on this factor. The alpha for the revised AS scale was .79 for both sexes. On the revised II scale, the alphas were .67 and .63 for men and women, respectively. The items on each scale are shown in Table 1.

Grade point average (GPA). The cumulative GPA for each subject was obtained from the students' official records following the 1986 spring semester. The number of semesters' work represented by this GPA varied from subject to subject, depending jointly on the semester (fall or spring) in which they were tested and the number of previous semesters

during which they had been enrolled at the University. The modal number of semesters' work represented in the GPA for the two samples as a whole was 2.

Health survey. A 32-item health survey was constructed whose items inquired about quality of sleep, problems of digestion and elimination, headaches, and respiratory problems (colds, flu, allergies). For subjects in the first sample, a priori scales in each of these four areas were constructed and whole-part correlations were determined for each item within each scale, separately for men and women. Ten items, the same for each sex, were dropped from further consideration on the basis of these analyses.

For the remaining items, a separate score was found for each subject on each content cluster. Scoring was such that high scores indicated better health. The alphas were all .75 or above in both sexes. In both sexes, correlations among the four health scores were all significantly positive ($p < .001$), with values ranging from .17 to .43. Therefore, an overall health measure, based on all 22 items, was obtained. The alphas for men and women were .82 and .83, respectively. Parallel analyses of the health data from the second sample provided very similar findings.

Other measures. Subjects in the first sample were given the three achievement motivation scales of the Work and Family Orientation Questionnaire (WOFO; Helmreich & Spence, 1978; Spence & Helmreich, 1978). These scales are labeled *Mastery* (preference for difficult challenging tasks), *Work* (the desire to work hard), and *Interpersonal Competitiveness* (the desire to compete against others and to win). The items are each accompanied by a 5-point rating scale.

Another personality instrument was also administered to subjects in Sample 2 for another purpose, the results of which will not be reported here.

Procedure

Subjects were tested in university classrooms in mixed sex groups of 80–100. They were not requested to supply their names but did provide their social security numbers so that it was possible to obtain their GPA from university records.

Results

Sample 1

Correlations within JAS. The correlations between the Jenkins et al. (1971) A–B, Factor H, and Factor S scales and the two

Table 2

Sample 1: Correlations Between the Three Original JAS Scales and the Two New JAS Scales.

JAS scale	Original JAS scale			New JAS scale	
	A-B	H	S	AS	II
A-B	—	.73	.80	.76	.43
H	.68	—	.58	.83	.37
S	.78	.53	—	.57	.70
AS	.73	.82	.52	—	.21
II	.40	.30	.67	.13	—

Note. $r_{.05} = .09$; $r_{.01} = .13$. JAS = Jenkins Activity Survey A-B = Type A-Type B scale; H = Factor H (Hard-driving competitiveness) scale; S = Factor S (speed and impatience) scale; AS = Achievement Strivings scale; and II = Impatience-Irritability scale. Correlations for men ($N = 362$) are above the diagonal and for women ($N = 351$) below the diagonal.

new factor scales, AS and II, are reported in Table 2. Although Jenkins et al. described the Factor H and S scales as independent, inspection of the correlations obtained with the three Jenkins scales shows that in each sex they are all substantially related to each other, in part because of overlapping content. They are also substantially related to the AS and II scales for the same reason. The AS and II scales, in contrast, are only modestly correlated.

JAS correlations with GPA and health. The correlations of the AS and the II scales with GPA and the health measures are shown in Table 3. The correlations between the health measures and the Jenkins et al. (1971) A-B scale are also reported. (Predictably, the relations of GPA and health with Factors H and S were similar to those with A-B; therefore, they will not be reported.)

As expected, significantly positive correlations were found between GPA and the AS scale in both sexes. The correlations with the II scale, on the other hand, were not only lower in magnitude but also negative in sign.

Also as anticipated, the opposite pattern occurred with the health measures. Negative relations with II occurred in both sexes on the several indexes (i.e., impatient, irritable men and women reported more health problems than their more placid peers). All r s were significant ($p < .05$) except for the digestion/elimination category in women. (See, however, the Sample 2 results in Table 5.) The correlations between AS scores and the health measures were all lower and, with the exception of sleep quality, nonsignificant.

Inspection of the data from the Jenkins et al. (1971) A-B measure shows, ironically, that the scale is a better predictor of GPA than of health. This outcome was not surprising in view of the stronger correlation between A-B and AS than between A-B and II and reflected the substantial presence of items related to achievement on the A-B scale.

Correlations with achievement motive scales. The correlations of the three WOFO achievement motive scales with the JAS scales, GPA, and the overall health measure are reported in Table 4. In both sexes, WOFO Mastery and Work scores were moderately correlated with the JAS A-B and AS scales. Only trivial relations were found with the II scale.

The relations with the WOFO Interpersonal Competitiveness scale are of particular interest because of the prominence of competitiveness in descriptions of the Type A pattern. This scale was moderately correlated with both the AS and II scales and with the A-B scale, again in both sexes (see Table 4). In light of the differential relations between the health measure and the JAS AS and II factors, expectations about relations between health and the Competitiveness scale were not clear. As it turned out, the relation was nonsignificant in men and slightly but significantly negative in women; that is, more competitive women reported more symptoms. (The difference between the correlations in males and females, however, was not significant.)

The remaining relations reported in Table 4 also differed for men and women. In women, Mastery and Work motives may be observed to be uncorrelated with overall health ratings, thus paralleling the health findings with the AS scale. However, significant positive relations were found in men. Those scoring higher in Work and in Mastery tended to report better health. As determined by Fisher's Z test, the difference between the r s for men and women was significant ($p < .01$) in both instances. These findings may well reflect genuine sex differences. However, it also seems possible that, in both sexes, the relation between health and this pair of achievement motives may tend to be slightly positive, perhaps because of the relations of these motive scores with other desirable personality characteristics that influence health more directly.

Although the results are not central to the purposes of this investigation, several comments should be made about the correlations of the WOFO achievement motives with GPA. Previous studies (e.g., Helmreich et al., 1980; Spence & Helmreich, 1983) have found—with several different types of groups and

Table 3

Sample 1: Correlations Between the JAS AS, II, and A-B Scales and GPA and Health Measures

GPA and health measure	JAS scale		
	AS	II	A-B
GPA			
Men	.36	-.04	.25
Women	.33	-.09	.15
Sleep quality			
Men	-.13	-.25	-.15
Women	-.03	-.16	-.13
Headaches			
Men	-.07	-.20	-.03
Women	.05	-.18	-.04
Digestion/Elimination			
Men	-.04	-.21	-.06
Women	-.04	-.04	-.03
Respiratory			
Men	.08	-.18	.08
Women	-.04	-.12	-.08
Health total			
Men	-.06	-.31	-.07
Women	-.03	-.18	-.11

Note. $r_{.05} = .09$; $r_{.01} = .13$. JAS = Jenkins Activity Survey; AS = Achievement Strivings scale; II = Impatience-Irritability scale; A-B = Type A-Type B scale; and GPA = grade point average. High health scores indicate good health. N men = 362; N women = 351.

Table 4
Sample 1: Correlations of the WOFO Achievement Motives Scales With the JAS Scales, GPA, and Overall Health

	JAS scale				
WOFO scale	AS	II	A-B	GPA	Health
Men (<i>N</i> = 362)					
Mastery	.40	.09	.45	.20	.22
Work	.36	.03	.36	.15	.22
Competitiveness	.30	.26	.32	.09	.03
Women (<i>N</i> = 351)					
Mastery	.47	.07	.49	.07	-.02
Work	.43	.04	.39	.10	.00
Competitiveness	.25	.29	.33	-.02	-.11

Note. $r_{.05} = .09$; $r_{.01} = .13$. WOFO = Work and Family Orientation Questionnaire; JAS = Jenkins Activity Survey; AS = Achievement Strivings scale; II = Impatience-Irritability scale; A-B = Type A-Type B scale; and GPA = grade point average.

performance measures—that Mastery and Work motives contribute positively to performance. As these two motives increase in strength, however, competitiveness has an increasingly deleterious effect on performance.

Unlike the AS scale, which makes reference to academic work, the WOFO scales are more abstract, making no reference to the situations and tasks that activate these motives. We suspect that it is, in part, for this reason that the WOFO scales were less successful in predicting grades than the AS scale. It should also be noted that most of the students were freshmen and that their GPA was based on two semester's work. In previous studies (e.g., Spence & Helmreich, 1983), we have consistently found more substantial correlations with cumulative GPAs that reflect a greater number of semesters' work. Many disinterested students tested in their freshman year have dropped out, and those remaining have had an opportunity to identify their academic interests and talents and to select courses that satisfy them.

Sample 2

As stated earlier, in the second sample, subjects were not given the entire JAS but only the AS and II scales (in slightly revised form). The correlations between these scales were .32 for men and .14 for women ($ps < .01$). Correlations of the scales with GPA and the health measures are reported in Table 5. The results for both sexes basically replicated those found with the first sample. Thus, significantly negative rs occurred between II and the health measures, whereas for AS, rs were nonsignificant. On GPA, the opposite pattern occurred: AS was positively related and II was nonsignificantly related to academic performance.

Discussion

Our data lead to two major conclusions. First, psychometric analyses of responses to both the student and the adult forms

of the JAS indicate the presence of two relatively independent factors (Pred et al., in press). These factors—Achievement Strivings (AS) and Impatience and Irritability (II), have conceptual similarities to Factor H (hard-driving competitiveness) and Factor S (speed and impatience) as described by Jenkins et al. (1971). However, our factor analyses resulted in cleaner factors in terms of item content and allowed us to develop two factor scales that have no item overlap and have substantially lower correlations with each other than the Jenkins et al. H and S scales.

Second, the results obtained with our AS and II scales reveal different patterns of relations with indexes of performance excellence, on the one hand, and of health, on the other. The data from the student samples unambiguously show that the AS scale is significantly and positively related to GPA, whereas the II scale has no significant effect on performance. We obtained similar results in a reanalysis of the JAS data obtained by Matthews et al. (1980) from a sample of male academic psychologists (Helmreich, Spence, & Pred, in press). Significant correlations were found between the AS scale and two measures of attainment—number of publications and number of citations by others to published work—but the correlations with II were nonsignificant.

Conversely, the data from the student samples consistently indicated significant relations between II and measures of health, such that more irritable and impatient men and women reported a greater number of physical complaints. Smaller and usually nonsignificant relations were found between these measures and the AS scale.

Additional evidence supporting the two-component model was recently reported by Chidester (1986) in a dissertation investigating factors related to the performance of jet transport

Table 5
Sample 2: Correlations Between the Revised JAS Scales and GPA and Health Measures

GPA and health measure	JAS scale	
	AS	II
GPA		
Men	.33	-.03
Women	.27	.04
Sleep quality		
Men	-.01	-.20
Women	-.01	-.39
Headaches		
Men	-.10	-.20
Women	-.02	-.24
Digestion		
Men	-.12	-.12
Women	-.06	-.31
Respiratory		
Men	-.08	-.12
Women	-.04	-.26
Health total		
Men	-.09	-.23
Women	-.01	-.43

Note. JAS = Jenkins Activity Survey; AS = Achievement Strivings scale; II = Impatience-Irritability scale; and GPA = grade point average. N men = 256; N women = 225. $r_{.05} = .13$; $r_{.01} = .17$.

pilots. Significantly positive correlations were found between the AS scale and ratings of performance related to crew management. In contrast, the II scale was not only nonsignificantly related to this measure but was negatively correlated with observers' evaluations of technical proficiency in flying. Pilots scoring high on the II scale also reported poorer quality of sleep and other physical problems during layovers than did low-scoring pilots.

The results of these studies that demonstrate a relation between the II scale and reports of relatively minor physical problems cannot automatically be generalized to CHD and other cardiovascular disorders. However, when these data are considered in conjunction with studies (e.g., Williams et al., 1980) that show an association between CHD and variables such as hostility, they add weight to the conjecture that affective reactions and behaviors incorporated in the AHA! Syndrome are the critical elements within the Type A pattern relevant to these disorders and, perhaps, to other major health problems as well (Price, 1982).

An equally important implication of our results is that men and women who are hard-driving, achievement-oriented, and, often as a consequence, successful, may not have a greater risk of CHD and other physical ailments than less ambitious persons. This conclusion follows jointly from our findings that achievement strivings do not per se have a negative influence on health and that there is considerable independence between the two sets of attributes tapped by the AS and II scales.

The role of competitiveness merits special attention. It is popularly assumed that in order to be successful, particularly in business, one must possess competitiveness as a personal attribute. Competitiveness is often associated as well with other aspects of success-motivated behaviors. Descriptions of the Type A pattern, for example, typically bracket the hard-driving and competitive attributes together.

It will be recalled that the WOFO Interpersonal Competitiveness scale (but not the Mastery and Work scales) was significantly correlated with II as well as with AS. Furthermore, previous research (Spence & Helmreich, 1983) has suggested that high scores on the Competitiveness scale are often associated with performance of poorer quality, particularly when competitiveness is combined with high levels of work and mastery motives.

It should be noted that many situations are inherently competitive in the sense that the supply of resources is less than the demand for them. Applying for a job or a research grant, selling goods and services in a limited market, or taking part in an athletic event or other contests with winners and losers are all examples. It seems reasonable to assume that if they are to achieve their goals in such situations, people must be willing to enter into competition with others and to risk failure, or even to enjoy the challenge of competitive contests. The WOFO Interpersonal Competitiveness scale, however, assesses people's desire to best and win out over other people. Even when competitiveness is described as a desirable attribute, it seems to incorporate these interpersonal aspects. This kind of interpersonal competitiveness, however, has a hostile, aggressive tinge to it that may be responsible both for the correlation of the WOFO Competitiveness scale with the II scale and for its negative effects on quality of performance.

A final comment should be made about the JAS and its usefulness in health research. The concordance between individuals classified as Type A or B by means of the JAS and the SI has typically been found to be only moderate (e.g., Matthews, Krantz, Dembroski, & MacDougall, 1982). Furthermore, the evidence suggests that differences between Type As and Type Bs in vulnerability to CHD are more likely to occur when the classification is done by means of the SI rather than the JAS (Matthews & Haynes, 1986). The basis for these discrepancies is not necessarily the greater validity of a clinical interview than that of an objectively scored self-report measure, as some have intimated. What is notable about the JAS is the relative absence of items referring to irritability, anger, and hostility, the very elements that seem to bring about the CHD-Type A relation. Rather than abandoning the JAS or other similar self-report measures in favor of the SI, it would seem more profitable in future research to use an array of self-report instruments to pinpoint more exactly the relevant personality factors associated with CHD and other disorders. The results of such studies could, in turn, lead the way to the development of a more valid package of instruments to replace current measures, including the II scale that we have carved out of the JAS. Development of such assessment instruments depends jointly on bringing greater conceptual clarity and greater psychometric sophistication to this area of research.

References

- Chidester, T. R. (1986). *Mood, Sleep and Fatigue Effects in Flight Operations*. Unpublished doctoral dissertation, University of Texas at Austin.
- Diamond, E. L. (1982). The role of anger and hostility in essential hypertension and coronary heart disease. *Psychological Bulletin*, 92, 410-443.
- Friedman, M., & Rosenman, R. H. (1959). Association of specific overt behavior pattern with blood and cardiovascular findings. *Journal of the American Medical Association*, 169, 1286-1296.
- Friedman, M., & Rosenman, R. (1974). *Type-A behavior and your heart*. Greenwich, CT: Fawcett.
- Glass, D. C. (1977). *Behavior patterns, stress, and coronary disease*. Hillsdale, NJ: Erlbaum.
- Greer, S., & Morris, T. (1975). Psychological attributes of women who develop breast cancer: A controlled study. *Journal of Psychosomatic Research*, 2, 147-153.
- Haynes, S. G., Feinleib, M., Levine, S., Scotch, N., & Kannel, W. B. (1978). The relationship of psychosocial factors to coronary heart disease in the Framingham Study II: Prevalence of heart disease. *Journal of Epidemiology*, 107, 384-402.
- Helmreich, R. L., & Spence, J. T. (1978). The Work and Family Orientation Questionnaire: An objective instrument to assess components of achievement motivation and attitudes toward family and career. *JSAS Catalog of Selected Documents in Psychology*, 8, 35. (Ms. No. 1677)
- Helmreich, R. L., Spence, J. T., Beane, W. E., Lucker, G. W., & Matthews, K. A. (1980). Making it in academic psychology: Demographic and personality correlates of attainment. *Journal of Personality and Social Psychology*, 39, 896-908.
- Helmreich, R. L., Spence, J. T., & Pred, R. S. (in press). Making it without losing it: Type A, achievement motivation, and scientific attainment revisited. *Personality and Social Psychology Bulletin*.
- Jenkins, C. D., Zyzanski, S. J., & Rosenman, R. H. (1971). Progress toward validation of a computer scored test for the Type A coronary prone behavior pattern. *Psychosomatic Medicine*, 33, 193-202.

- Krantz, D. S., Glass, D. C., & Snyder, M. L. (1974). Helplessness, stress level, and the coronary prone behavior pattern. *Journal of Experimental Social Behavior*, 10, 284-300.
- Manuck, S. B., Morrison, R. L., Bellack, A. S., & Polefrone, J. M. (1985). Behavioral factors in hypertension: Cardiovascular responsivity, anger and social competence. In M. A. Chesney and R. H. Rosenman (Eds.), *Anger and hostility in cardiovascular and behavioral disorders* (pp. 149-172). Washington, DC: Hemisphere.
- Matthews, K. A. (1982). Psychological perspectives on the Type A behavior pattern. *Psychological Bulletin*, 91, 293-323.
- Matthews, K. A., Glass, D. C., Rosenman, R. H., & Bortner, R. W. (1977). Competitive drive, pattern A, and coronary heart disease: A further analysis of some data from the Western Collaborative Group Study. *Journal of Chronic Diseases*, 30, 489-498.
- Matthews, K. A., & Haynes, S. G. (1986). Type A behavior pattern and coronary disease risk. *Journal of Epidemiology*, 123, 923-960.
- Matthews, K. A., Helmreich, R. L., Beane, W. E., & Lucker, G. W. (1980). Pattern A, achievement striving, and scientific merit: Does pattern A help or hinder? *Journal of Personality and Social Psychology*, 39, 962-967.
- Matthews, K. A., Krantz, D. S., Dembroski, T. M., & MacDougall, J. M. (1982). Unique and common variance in Structured Interview and Jenkins Activity Survey measure of the type A behavior pattern. *Journal of Personality and Social Psychology*, 42, 303-313.
- Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. (1975). *SPSS: Statistical package for the social sciences* (2nd ed.). New York: McGraw-Hill.
- Pred, R. S., Helmreich, R. L., & Spence, J. T. (in press). The development of new scales for the Jenkins Activity Survey measure of the Type A construct. *Social and Behavioral Science Documents*.
- Price, V. A. (1982). *Type A behavior pattern: A model for research and practice*. New York: Academic Press.
- Rosenman, R. H. (1978). The Interview method of assessment of the coronary-prone behavior pattern. In T. Dembroski, S. M. Weiss, J. Shields, S. Haynes, & M. Feinleib (Eds.), *Coronary-prone behavior* (pp. 55-69). New York: Springer-Verlag.
- Rosenman, R. H. (1985). Health consequences of anger and implications for treatment. In M. A. Chesney & R. H. Rosenman (Eds.), *Anger and hostility in cardiovascular and behavioral disorders* (pp. 103-126). New York: Hemisphere/McGraw-Hill.
- Rosenman, R. H., Friedman, M., Straus, R., Wurm, M., Kositchek, R., Hahn, W., & Wethessen, N. T. (1964). A predictive study of coronary heart disease: The Western Collaborative Group Study. *Journal of the American Medical Association*, 189, 15-22.
- Spence, J. T., & Helmreich, R. L. (1978). *Masculinity and femininity: Their psychological dimensions, correlates and antecedents*. Austin: University of Texas Press.
- Spence, J. T., & Helmreich, R. L. (1983). Achievement-related motives and behavior. In J. T. Spence, (Ed.), *Achievement and achievement motives: Psychological and sociological approaches* (pp. 10-74). San Francisco: Freeman.
- Spielberger, C. D., Johnson, E. H., Russell, S. F., Crane, R. J., Jacobs, G. A., & Worden, T. J. (1985). The experience and expression of anger: Construction and validation of an anger expression scale. In M. A. Chesney and R. H. Rosenman (Eds.), *Anger and hostility in cardiovascular and behavioral disorders* (pp. 5-30). New York: Hemisphere/McGraw-Hill.
- Taylor, M. S., Locke, E. A., Lee, C., & Gist, M. (1984). Type A behavior and faculty research productivity: What are the mechanisms? *Organizational Behavior and Human Performance*, 34, 402-418.
- Waldron, I., Hickey, A., McPherson, C., Butensky, A., Gruss, L., Overall, K., Schmader, A., & Wohlmuth, D. (1980). Type A behavior pattern: Relationship to variation in blood pressure, parental characteristics, and academic and social activities of students. *Journal of Human Stress*, 6, 16-27.
- Weinstein, K. A., Davison, G. C., DeQuattro, V., & Allen, J. W. (1986, August). *Type A behavior and cognitions: Is hostility the bad actor?* Paper presented at the 94th Annual Convention of the American Psychological Association, Washington, DC.
- Williams, R. B., Haney, T. L., Lee, K. L., Kong, Y., Blumenthal, J. A., & Whalen, R. E. (1980). Type A behavior, hostility, and atherosclerosis. *Psychosomatic Medicine*, 42, 539-549.

Received October 20, 1986

Revision received February 6, 1987

Accepted February 11, 1987 ■

Call for Nominations for *Journal of Experimental Psychology: General*

The Publications and Communications Board has opened nominations for the editorship of the *Journal of Experimental Psychology: General* for the years 1990-1995. Sam Glucksberg is the incumbent editor. Candidates must be members of APA and should be available to start receiving manuscripts in early 1989 to prepare for issues published in 1990. Please note that the P & C Board encourages more participation by women and ethnic minority men and women in the publication process, and would particularly welcome such nominees. To nominate candidates, prepare a statement of one page or less in support of each candidate. Submit nominations no later than February 15, 1988 to

Donald J. Foss
Department of Psychology
University of Texas
Austin, Texas 78712

Other members of the search committee are James J. Jenkins, Jean Mandler, J. E. R. Staddon, and Saul Sternberg.

Occupational Complexity, Control, and Personal Income: Their Relation to Psychological Well-Being in Men and Women

Pamela K. Adelman
University of Michigan

Research on work and well-being indicates that paid employment has beneficial consequences for mental health. In this study, it was hypothesized that higher occupational complexity, control, and personal income would be associated with higher levels of happiness and self-confidence and lower psychological vulnerability. In addition, the possibility was explored that models describing these correlations for employed women ($n = 330$) might differ from those for employed men ($n = 618$). Stepwise regression results indicate that occupational characteristics explain a small but significant proportion of variance in each measure of psychological well-being controlling for the effects of age and education. In addition, analysis of covariance reveals that separate regressions characterize employed men and women for happiness and self-confidence but not for vulnerability. Occupational characteristics also explain a significant proportion of variance in self-confidence for both men and women, and in happiness for men. Implications for the relation of work to well-being and for job enrichment and enlargement programs are discussed.

Increasing evidence suggests that paid employment can have beneficial consequences for psychological well-being. Employed men have higher well-being than do men who are not employed (Cobb & Kasl, 1977; Veroff, Douvan, & Kulka, 1981), and paid employment also has a positive relation to psychological well-being in women (see Adelman, 1987, or Warr & Parry, 1982, for a review of this extensive literature), although there are some exceptions to this finding (Campbell, 1980; Warr & Parry, 1982; Wright, 1978). Yet what it is about paid employment that benefits well-being remains relatively unexplored. The purpose of this study was to test whether three occupational characteristics (control, complexity, and personal income) are related to psychological well-being in employed men and women.

A substantial literature exists on the relation between occupational characteristics and one psychological outcome, job satisfaction. For example, it is well established that income is associated with higher job satisfaction (Kahn, 1972; Lawler, 1971; Martin & Hanson, 1986). Control (variously conceptualized as autonomy, responsibility, decision latitude, and supervisory status) has also been found to have a positive association with job satisfaction (Hackman, Pearce, & Wolfe, 1978; Karasek, 1979; Porter & Lawler, 1965; Turner & Lawrence, 1965; Vroom, 1964; Weaver, 1977), and occupational complexity (variety, skill level, task significance, lack of repetition or routinization, and challenge) has a similar relation to job satisfaction (Eichar & Thompson, 1986; Hackman et al., 1978; Hackman & Lawler, 1971; Kohn & Schooler, 1973; Turner & Lawrence, 1965).

However, job satisfaction represents a specific facet of psychological well-being and one that is especially likely to have an association with job characteristics. Relatively few studies have examined the relation between job characteristics and other, broader psychological outcomes.

Pay (Near, Rice, & Hunt, 1978; Near, Smith, Rice, & Hunt, 1983), control (Karasek, 1979), and complexity (Caplan, Cobb, French, Harrison, & Pinneau, 1975; Gardell, 1971) are related to outcomes reflecting happiness or general positive affect, such as life satisfaction and morale. Some characteristics are also associated with reduction in negative psychological reactions. For example, occupational control is associated with lower depression (Karasek, 1979) and anxiety (Kohn, 1969). Complexity is also associated with lower anxiety (Kohn, 1969; Kohn & Schooler, 1973; Kornhauser, 1965; Miller, Schooler, Kohn, & Miller, 1979) and neurotic illness (Kornhauser, 1965). Occupational characteristics are also related to various facets of self-evaluation, such as self-esteem, self-confidence, self-deprecation, and job self-esteem. Higher control is associated with more positive self-evaluation (Hackman et al., 1978; Kohn, 1969), as is greater occupational complexity (Caplan et al., 1975; Gardell, 1971; Hackman & Lawler, 1971; Kohn, 1969; Kohn & Schooler, 1973; Kornhauser, 1965; Miller et al., 1979).

There are two major weaknesses in the existing literature on occupational characteristics and global psychological well-being. Many of these studies focus on narrow segments of the labor force, such as factory or assembly-line workers, rather than on the full occupational spectrum. In addition, women are frequently excluded from these studies, and when they are included, sex differences are rarely tested (Haw, 1982).

In summary, the existing research on the relation between occupational characteristics and psychological well-being is limited in a number of ways. In this article, the relation between occupational characteristics and psychological well-being was more fully explored. A broad set of measures of psychological

The author wishes to thank Joseph Veroff, Elizabeth Douvan, and Richard Kulka for the use of their data, and Joseph Veroff, Lerita Coleman, Toni Antonucci, and the anonymous reviewers for their valuable comments on earlier versions of this article.

Correspondence concerning this article should be addressed to Pamela K. Adelman, Institute for Social Research, University of Michigan, Ann Arbor, Michigan 48106-1248.

well-being was examined; the sample used was nationally representative and spanned the range of occupational categories; and the possibility that different models characterize the relation between occupational characteristics and well-being in men and women was tested.

In addition, an important feature of this study was that it used independent ratings of occupational characteristics made by trained observers rather than self-reports about jobs. The use of outside evaluations minimizes the possibility that employees' perceptions of job characteristics may be influenced by job satisfaction (Caldwell & O'Reilly, 1982; O'Reilly & Caldwell, 1979), life satisfaction (Keon & McDonald, 1982), or other dimensions of well-being.

The central hypothesis of this study was that higher occupational complexity, control, and personal income are positively related to psychological well-being (happiness, self-confidence, and lack of psychological vulnerability to negative experiences). Because so few previous studies on the effects of occupational conditions have included women or tested for sex effects, no specific hypotheses regarding possible sex differences were formulated.

Method

Data

Data were collected by face-to-face interview in a 1976 cross-sectional national survey of adults age 21 or over (Veroff et al., 1981). Although these data are now more than 10 years old, the breadth of this survey has not since been matched in its attempt to assess the mental health of a national cross section of adults. It is also unusual because several of the ratings of job characteristics from the *Dictionary of Occupational Titles* (DOT; U.S. Department of Labor, 1965) are merged with the survey's data on occupations. Thus, this data set represents the most recent and richest source of information for the purposes of this study.

Variables

Occupational characteristics. Eight measures of occupational characteristics were available from this data set. They include seven variables from the DOT and a self-report measure of personal earnings.

Personal income was assessed by a self-report item measuring income personally earned by the respondent through paid labor. It is an 18-level ordinal variable, with the first 11 levels in \$1,000 increments from \$0–\$999 to \$10,000–\$10,999, and the last seven levels in larger increments (\$11,000–\$12,499, \$12,500–\$14,999, \$15,000–\$17,499, \$17,500–\$19,999, \$20,000–\$24,999, \$25,000–\$34,999, and \$35,000 or over).

The variables from the DOT are aggregated at the occupational level. Ratings were made for the 21,749 detailed occupations in the DOT by trained job analysts at the Department of Labor. These ratings were merged into the U.S. Census occupational categorization—which contains only 584 titles—by Temme (1975), and thus could be combined with the current data set, which uses the census classification. In merging the two classification systems, Temme calculated weighted means of the characteristics according to the proportions represented by each of the multiple categories of the DOT fitting into each census category.

Four of the DOT variables are from the Temperaments section and are described as “types of occupational situations to which workers must adjust” (U.S. Department of Labor, 1965, p. 654). *Variety* indicates the degree to which an occupation involves a variety of duties characterized by frequent change; *repetition* reflects the extent to which fixed, short-cycle operations are involved. *Instruction by others* involves

Table 1
Factor Loadings of Occupational Characteristics^a

Item	Factor 1	Factor 2
Variety	.75	–.19
Repetition	–.89	–.22
Instruction by others	–.79	–.28
Involvement with data	.73	.50
Involvement with people	.29	.83
Involvement with things	.15	–.81
Control, self or others	.42	.63
Percentage common variance	40.1	30.9

Note. N = 932.
^a Principal components analysis with varimax rotation and Kaiser criterion of eigenvalue 1. Kaiser's statistic = .84.

doing work under the close supervision of others, with little allowance for independent action or judgment. *Control over self or others* measures the degree of planning and control of one's own activities and those of others.

Involvement with data includes the use of information, knowledge, and conceptions in an occupation; the highest levels include synthesizing and coordinating data, whereas the lowest levels include copying and comparing data. *Involvement with people* describes the nature of the worker's relations with others at work and ranges from mentoring, instructing, and negotiating with others to serving other people. Finally, *involvement with things* includes contact with machines, tools, and products and ranges from setting up and working equipment to simply handling objects with little judgment required. (See Appendix A for detailed definitions of the DOT variables.)

The four temperament variables were originally coded as either present or absent to a significant degree for each occupation. Scores on involvement with data, people, and things ranged from 0 to 8, with low scores reflecting greater involvement (scores were reversed for easier interpretation in this study).

Factor Analysis of Occupational Characteristics

Because of high statistical and conceptual interrelatedness among the seven DOT items, a factor analysis was conducted on these variables. Exploratory factor analysis (principal components analysis with Kaiser criterion set to eigenvalue 1) using orthogonal rotation resulted in two uncorrelated factors (see Table 1).

Variety and involvement with data have high positive loadings (>.40) on Factor 1; repetition and instruction by others both load negatively (<–.40). Each of these four items touches on the complexity of the work task itself. High variety and low repetition are considered to constitute structural complexity (Kohn, 1969), and greater involvement with data reflects higher substantive complexity of work (Kohn, 1969); instruction by others involves the degree of “action or judgment in working out job problems” (see Appendix A), and a low score on this variable thus also seems to reflect greater occupational complexity. Factor 1 was therefore labeled *Complexity*. This factor accounted for 40.1% of the common variance. (Control over self and others also loaded at >.40 on this factor, but loaded more highly on Factor 2.)

The second factor accounts for an additional 30.9% of the common variance. The items loading highly and positively on this factor are involvement with people and control over self and others (involvement with data, despite a loading >.40, loaded more highly on Factor 1); involvement with things loaded negatively (<–.40). This factor was named *Control* because the items reflect the latitude one has in approaching the work task. Control over self and others involves control-

ling and planning one's own and others' work; involvement with people implies a continuum of degree of control over others (from mentoring, negotiating, and instructing, at the highest levels, to serving others, at the lowest). Similar constructs have been included in definitions of occupational control (e.g., Hackman & Lawler, 1971; Kahn, 1981). Involvement with things may reflect the degree to which one's work is controlled by machines and objects, also considered a facet of autonomy and control (Hackman & Lawler, 1971; Kahn, 1981).

Standardized factor scores ($M = 0$, $SD = 1$) were created for use in subsequent analyses.

Psychological well-being. The three measures chosen to represent the positive, self-evaluative, and negative aspects of psychological well-being were happiness, self-confidence, and vulnerability. These variables were among the factors extracted in a factor analysis of 25 measures in the present data set performed by Bryant and Veroff (1984). The confirmatory factor analysis they conducted revealed that a single model of well-being fit both men and women equally well. The factor loadings on the three factors used here appear in Appendix B; further details of the analyses are available in Bryant and Veroff (1984).

Happiness is described as a global, spontaneous, affective evaluation of life in general. The items loading on this factor are all oriented toward positive experience and include degree of general happiness, whether present is happier than the past, whether one's happiest time is in the present, degree of morale regarding the future, and degree of satisfaction with life in general. A high score indicates greater happiness.

The self-confidence factor reflects distress and well-being in the specific domain of the self. It incorporates degree of depression, anomie, self-esteem and self-acceptance, and the perception of controllability of problems and outcomes. Bryant and Veroff (1984) suggested that this factor reflects people's evaluations of their ability to "effect positive experiences in their lives and to handle negative events that occur" (p. 17). A high score reflects greater self-confidence.

Vulnerability, in contrast to happiness, focuses on general, spontaneous evaluations of negative experiences. The items included in this factor are feelings of being overwhelmed, having ever felt about to have a nervous breakdown, and perceptions of bad things occurring frequently. Higher scores indicate greater vulnerability.

Three other factors were not included in the analyses. Gratification focuses on satisfaction and value fulfillment in specific life roles—job, marriage, parenthood, leisure, and housework. It seemed unlikely that the specific attributes of only one of these roles (job) would have a substantial relation with the entire factor. The strain factor includes psychophysical symptoms such as anxiety, ill health, immobilization, and drug and alcohol abuse, which are likely to be associated with stressful characteristics of work not included here, such as time pressure, noise and quantitative demands (Frese, 1985; Kahn, 1972; Ritti, 1971), and physical working conditions. The last factor, uncertainty, was considered something of a "residual factor" by Bryant and Veroff (1984) and, in addition, was the only factor on which men and women substantially differed. Its items include frequency of worrying, economic worries, admission of shortcomings in the self, and dissatisfaction with personal time use. For these reasons, these three factors were omitted from this study.

Bryant and Veroff (1984) noted considerable overlap among the measures comprising the six factors (oblique rotation was used). They therefore suggested that these factors should best be considered multiple indicators of psychological well-being rather than sharply distinct factors.

Background variables. Age and education were the two demographic characteristics found to be most important to the understanding of subjective mental health in the detailed analysis of subjective well-being by Veroff et al. (1981). Similarly, age and education were found to correlate with a number of the well-being factors previously described (Bryant & Veroff, 1984). In addition, older and more highly educated employees are likely to have higher income and greater occupational complexity

and control. The effects of these two variables were therefore controlled in these analyses.¹ The age variable was intervally scaled, and education was ordinal with eight levels: *less than one grade completed* (1), *1–6 grades* (2), *7–8 grades* (3), *9–11 grades* (4), *12th grade or high school graduate* (5), *some post-high school training* (6), and *some college* (7), *college graduate* (8). Both age and education were assessed via self-report items.

Sample Characteristics

The sample of 2,264 adults was drawn using area-sampling probability methods, as described in detail in Kish and Hess (1965). Within this representative sample, 618 men between the ages of 21 and 65 years were employed full time (40 or more hours per week) and 330 women in this age range were employed full time. Of the full sample of 2,264 adults, 1,316 respondents were excluded—909 did not work any hours for pay (504 homemakers, 229 retired, 84 unemployed, 52 disabled, 20 laid off, and 20 students), 404 worked part time (1–39 hr) or were employed but were over 65 years old, and 3 cases had missing data.

Means and standard deviations for the occupational characteristics, well-being, and background variables appear in Table 2. Men and women did not significantly differ on age or educational level; their average age was roughly 38 to 39 years, and the mean level of education for both groups was at the high school graduate level.

However, employed men earned significantly higher incomes than did employed women ($t = 18.00$, $p < .01$) and had more occupational complexity ($t = 2.42$, $p < .05$), but less occupational control ($t = -2.24$, $p < .05$).

Men and women did not significantly differ in mean levels of happiness and self-confidence. Women had significantly higher vulnerability than did men ($t = -7.06$, $p < .01$).

Data Analyses

Three sets of analyses were performed to test the hypotheses of this study. First, stepwise regressions for happiness, self-confidence, and vulnerability were conducted to test the hypothesis that greater income, complexity, and control are related to higher psychological well-being net of the influence of age and education. In the first step, the background variables of age and education were entered as one block. In the second step, income, complexity, and control were entered and tested for an incremental increase in explained variance.

Next, to explore the possibility that different regression models describe the relation of occupational complexity, control, and income with well-being for men and women, a set of analyses of covariance (ANCOVAs) was performed. In these analyses, the null hypothesis was tested that regression slopes of well-being on age, education, income, complexity, and control do not differ for men and women.

Finally, for those well-being measures in which the regression slopes were found to differ by sex, separate stepwise regression analyses were conducted for men and for women.

Results

Combined Regression Analyses

The correlation matrix of background, occupational, and well-being variables used in the regression analyses appears in

¹ Subsequent examination of the sample correlation matrices revealed that age is not significantly correlated with occupational complexity or control, contrary to expectations. This suggests that in future research on occupational characteristics and well-being, its inclusion as a control variable may be unnecessary.

Table 2
Mean Levels of Background Variables, Occupational Characteristics, and Psychological Well-Being

Variable	Men and women (n = 948)		Men (n = 618)		Women (n = 330)		t
	M	SD	M	SD	M	SD	
Age	38.67	12.12	39.01	11.94	38.03	12.45	1.19
Education	5.93	1.60	5.95	1.65	5.88	1.51	0.72
Income	10.93	4.18	12.49	3.63	7.99	3.52	18.00**
Complexity	0.00	1.00	0.06	1.06	-0.11	0.88	2.42*
Control	0.00	1.00	-0.05	1.08	0.10	0.84	-2.24*
Happiness	29.23	6.37	29.30	6.22	29.10	6.66	0.47
Self-confidence	33.62	5.44	33.85	5.13	33.18	5.95	1.80
Vulnerability	22.12	7.06	20.96	6.20	24.28	8.01	-7.06**

* $p < .05$. ** $p < .01$.

Table 3. Various indicators of multicollinearity are acceptably low for this sample. First, the zero-order correlations among predictors within each block are moderately low. For example, the correlation between age and education is less than $-.20$, and the highest intercorrelation among the occupational characteristics is only $.22$. Second, the determinant of the matrix (which can range from 0 to 1, where 1 indicates perfect multicollinearity) is an acceptable $.359$. Finally, the regression program used computes a covariance ratio for each predictor; this figure shows the proportion of variance in each predictor explained by regressing it on all other predictors in the equation. Generally, a covariance ratio of $<.70$ is considered acceptable; in these regressions, the highest covariance ratio for any predictor is $.58$.

The results of the regressions for men and women combined are contained in Table 4. The partial correlations (*partial r*) and regression coefficients (β) are reported for each variable at the step in which it was entered in a block. The incremental and cumulative multiple correlations are reported at each step (they are identical at the first step).

In the first step of the regressions, happiness, self-confidence, and vulnerability were each regressed on the background variables of age and education. For each dependent measure a significant proportion of variance is explained by the background variables. These proportions are 2.2% for happiness, $F(2, 882) = 9.74, p < .01$; 9.2% for self-confidence, $F(2, 882) =$

44.77, $p < .01$; and 0.7% for vulnerability, $F(2, 882) = 3.05, p = .05$.

Significant regression coefficients indicate that older respondents are significantly less happy than younger ones, $t(1, 882) = 2.75, p < .01$, but more self-confident, $t(1, 882) = 3.13, p < .01$, and less vulnerable, $t(1, 882) = 2.41, p < .05$. Higher educational attainment is associated with greater happiness, $t(1, 882) = 2.90, p < .01$, and self-confidence, $t(1, 882) = 9.35, p < .01$.

The second step of the regressions shows that the block containing occupational characteristics explains a significant proportion of variance in well-being after age and education have been entered. The incremental proportion of variance in happiness explained by occupational characteristics is 2.3%, $F(3, 879) = 7.17, p < .01$; the cumulative proportion of variance is 4.5%, $F(5, 879) = 8.28, p < .01$. For self-confidence, an additional 3.7% of variance is explained by the addition of occupational characteristics, $F(3, 879) = 12.55, p < .01$, for a cumulative proportion of variance of 12.9%, $F(5, 879) = 26.14, p < .01$. Finally, 4% of the variance in vulnerability is explained by occupational characteristics net of the effects of background variables, $F(3, 879) = 12.41, p < .01$. They bring the total variance explained to 4.7%, $F(5, 879) = 8.71, p < .01$.

Income bears a significant relation to all measures of mental health. Higher income is related to greater happiness, $t(1,$

Table 3
Correlations Among Background Variables, Predictors, and Dependent Variables

Variable	1	2	3	4	5	6	7	8
1. Age	—							
2. Education	-.18*	—						
3. Income	.11*	.37*	—					
4. Complexity	-.01	.23*	.20*	—				
5. Control	.04	.45*	.22*	.01	—			
6. Happiness	-.11*	.11*	.15*	.07	.10*	—		
7. Self-confidence	.05	.29*	.26*	.12*	.24*	.38*	—	
8. Vulnerability	-.08	-.02	-.20*	-.04	-.02	-.17*	-.25*	—

Note. $N = 85$.
* $p < .01$.

Table 4
Regression Analyses of Happiness, Self-Confidence, and Vulnerability

Predictor	Happiness			Self-Confidence			Vulnerability		
	Partial <i>r</i>	β	R^2	Partial <i>r</i>	β	R^2	Partial <i>r</i>	β	R^2
Background variables ^a									
Age	-.09	-.09**		.10	.10**		-.08	-.08*	
Education	.10	.10**		.30	.30**		-.03	-.03	
R^2 at this step			.022 ^b			.092 ^b			.007 ^b
Occupational characteristics ^a									
Income	.13	.14**		.15	.15**		-.19	-.21**	
Complexity	.04	.04		.05	.05		-.00	-.00	
Control	.07	.08*		.12	.13**		.05	.05	
Incremental R^2			.023 ^b			.037 ^b			.040 ^b
Cumulative R^2			.045 ^b			.129 ^b			.047 ^b

^a Standardized regression coefficients and partial correlations are reported for variables at this step in the regression.

^b Variables in this block add significantly to the explanatory power of the model at $p \leq .05$.

* $p < .05$. ** $p < .01$.

879) = 3.74, $p < .01$; more self-confidence, $t(1, 879) = 4.42$, $p < .01$; and lower vulnerability, $t(1, 879) = 5.88$, $p < .01$. In contrast, complexity is not significantly associated with any of the three dependent measures. Greater occupational control is correlated with increased happiness, $t(1, 879) = 2.15$, $p < .05$, and self-confidence, $t(1, 879) = 3.67$, $p < .01$.

To summarize, in this set of regression analyses, the background variables of age and education as a block explain a significant proportion of variance in happiness, self-confidence, and vulnerability. The occupational characteristics explain a significant additional proportion of variance in all three measures of well-being.

Analyses of Covariance

An ANCOVA was performed for each outcome variable to test the hypothesis that the regression equations of well-being on age, education, income, control, and complexity have the same slopes for men and women. The results of these analyses are reported in Table 5.

The test of interest for the purposes of this study is the test of equal slopes. The analyses reveal that for happiness, $F(5, 873) = 95.22$, $p < .05$, and self-confidence, $F(5, 873) = 56.54$, $p < .05$, the null hypothesis is rejected. This suggests that significantly different regression equations characterize the relation between occupational characteristics and well-being for men and women, and that separate regression analyses within each group are appropriate on these variables. On vulnerability, however, the slopes for men and women do not significantly differ; the hypothesis that the slopes are the same for men and women cannot be rejected. Therefore, the regression equation for vulnerability was not tested separately within each group.

Separate Regression Analyses

The correlation matrices used in regressions run separately for men and women appear in Table 6. Correlations for men

appear above the diagonal, and those for women appear below the diagonal.

Table 7 contains the results of separate regression analyses for men and women for happiness and self-confidence. The stepwise procedures described earlier were repeated within each group. Indicators of multicollinearity are again within acceptable limits; the determinant of the matrix for women is .409, and for men it is .576. The highest covariance ratio for women is .45, and for men is .35.

Table 5
Analysis of Covariance for Happiness, Self-Confidence, and Vulnerability

Source	<i>df</i>	<i>MS</i>	<i>F</i>
Happiness			
Equal regressions	6	103.20	2.72*
Equal adjusted means	1	143.09	3.77*
Equal slopes	5	95.22	2.51*
Error	873	37.94	
Total	884		
Self-Confidence			
Equal regressions	6	50.26	1.95
Equal adjusted means	1	18.90	0.73
Equal slopes	5	56.54	2.19*
Error	873	25.79	
Total	884		
Vulnerability			
Equal regressions	6	177.37	3.84**
Equal adjusted means	1	924.37	20.02**
Equal slopes	5	27.97	0.61
Error	873	46.17	
Total	884		

* $p < .05$. ** $p < .01$.

Table 6
Correlations Among Background Variables, Predictors, and Dependent Variables for Men and Women

Variable	1	2	3	4	5	6	7
1. Age	—	-.15*	.15*	.04	.07	-.04	.06
2. Education	-.24*	—	.39*	.16*	.44*	.10	.26*
3. Income	.00	.51*	—	.17*	.25*	.20*	.29*
4. Complexity	-.13	.38*	.22*	—	-.04	.02	.05
5. Control	-.03	.48*	.45*	.16*	—	.13*	.24*
6. Happiness	-.23*	.15*	.11	.16*	.05	—	.39*
7. Self-confidence	.01	.33*	.25*	.24*	.27*	.37*	—

Note. $N = 883$ (571 men, 312 women). Correlations above the diagonal are for men, below the diagonal for women.
* $p < .01$.

For men and women, background variables as a block account for a significant proportion of variance in both happiness and self-confidence. They explain 6.1% of the variance in happiness among women, $F(2, 309) = 10.07, p < .01$, and 1.1% among men, $F(2, 570) = 3.02, p = .05$. The proportion of variance explained by the block for women is significantly greater than the amount explained for men ($z = 2.08, p < .05$, two-tailed). Older women are significantly less happy than younger ones, $t(1, 309) = 3.63, p < .01$, and men with more education are happier than less educated men, $t(1, 570) = 2.27, p < .05$. Age and education explain 12.2% of the variance in self-confidence for women and 8% of the variance for men (these proportions do not significantly differ). Significant beta coefficients indicate that a higher level of education is associated with greater self-confidence in both men, $t(1, 570) = 6.86, p < .01$, and women, $t(1, 309) = 6.54, p < .01$. Age is also significantly correlated with self-confidence in men, $t(1, 579) = 2.62, p < .01$; older men are more self-confident. The block of occupational characteristics does not add sig-

nificantly to the proportion of variance in happiness explained by the background characteristics among women. The incremental proportion of variance is 1.6%, $F(3, 306) = 1.73, p = .16$, and the cumulative variance increases to only 7.7%. Occupational characteristics do add significantly to the explained variance in happiness among men; an additional 4.2% of variance is accounted for, $F(3, 567) = 8.53, p < .01$, and the cumulative variance rises to 5.3%, $F(5, 567) = 6.37, p < .01$. The incremental and cumulative proportions of variance explained in happiness are not significantly different for men and women. For both men and women, occupational characteristics add significantly to the variance in self-confidence. They explain 3.1% for women, $F(3, 306) = 3.72, p < .05$, and 4.8% for men, $F(3, 567) = 10.34, p < .01$, over and above the effects of background variables. Total variance explained for women is 15.3%, $F(5, 306) = 11.02, p < .01$, and for men is 12.8%, $F(5, 567) = 16.62, p < .01$. Men and women do not significantly differ in the incremental or cumulative proportions of variance explained in self-confidence.

Table 7
Separate Regression Analyses of Happiness and Self-Confidence for Men and Women

Predictor	Happiness						Self-confidence					
	Women			Men			Women			Men		
	Partial r	β	R^2	Partial r	β	R^2	Partial r	β	R^2	Partial r	β	R^2
Background variables ^a												
Age	-.20	-.21**		-.03	-.03		.10	.10		.11	.11**	
Education	.10	.10		.09	.09*		.35	.36**		.28	.28**	
R^2 at this step			.061 ^b			.011 ^b			.122 ^b			.080 ^b
Occupational characteristics ^a												
Income	.07	.08		.18	.20**		.05	.05		.18	.20**	
Complexity	.10	.11		.00	.00		.13	.13*		.00	.00	
Control	-.02	-.02		.09	.11*		.11	.12*		.12	.13**	
Incremental R^2			.016			.042 ^b			.031 ^b			.048 ^b
Cumulative R^2			.077 ^b			.053 ^b			.153 ^b			.128 ^b

^a Standardized regression coefficients and partial correlations are reported for variables at this step in the regression.
^b Variables in this block add significantly to the explanatory power of the model at $p \leq .05$.
* $p < .05$. ** $p < .01$.

In summary, separate regressions of happiness and self-confidence reveal that background variables contribute significantly, though slightly, to explained variance for both men and women. Occupational characteristics explain some additional variance in happiness for men, and in self-confidence for employed men and women.

Discussion

The impact of employment status on well-being has been extensively documented by research conducted during the past 20 years, but what aspects of employment are important to well-being remains relatively unexplored. The present study investigated the hypothesis that certain occupational characteristics (personal income, complexity, and control) are related to psychological well-being in employed men and women even after the effects of age and education are controlled. The results suggest that as a block, these variables are significantly related to well-being, but that the pattern of results varies somewhat for men compared with women and for each of the three characteristics.

A single regression equation characterized the relation between occupational characteristics and vulnerability for women and men. The background variable of age was positively associated with vulnerability to negative events. Of the occupational characteristics, higher income was associated with lower vulnerability. Neither occupational complexity nor control was significantly related to vulnerability.

For happiness and self-confidence, ANCOVA revealed that the regression equations for men and women significantly differ. More education, higher income, and greater control were related to increased happiness in men, and increasing age was related to lower happiness among women. None of the three occupational characteristics was significantly associated with happiness in employed women, and it may be that experiences in other roles (such as marriage and parenthood) may be more important to women's happiness than are occupational characteristics.

For both women and men, the background variable of education was associated with greater self-confidence; age was also related to higher self-confidence in men. Occupational control was significantly related to the self-confidence of both men and women. In addition, higher income in men and greater occupational complexity among women were associated with increased self-confidence.

The proportion of variance explained by occupational characteristics over and above the effects of age and education was significant for every equation except happiness for women. Yet the magnitude of these proportions was quite small—less than 5% in all equations. In addition, only in the vulnerability regression and the happiness regression for men did the proportion of variance accounted for by the block of occupational characteristics exceed the proportion explained by the background variables.

In addition, less variance was explained by occupational characteristics among employed women than men for both self-confidence and happiness (although the differences in proportions were not significant). A restricted range of occupational complexity and control among women (suggested by their

smaller standard deviations on both variables) may account for this. Because there is less variability in women's occupational characteristics, they may be less capable of explaining variance in psychological well-being among women.

The relation between occupational characteristics and psychological well-being is probably also underestimated in this study because only three characteristics were included. Rice, Near, and Hunt (1979) found that "general life-satisfaction can be influenced by work-related variables," and added that because only two such variables were included in their study, "this conclusion is probably conservative" (p. 618). Other occupational characteristics, such as physical demands, affiliative contacts with coworkers, time pressures, temperature and cleanliness—variables that were not available in the present data set—would probably add to the explained variance in well-being if incorporated into the model.

One question that has frequently arisen in the context of research on the relation of job characteristics to job satisfaction is the direction of causality. This question is also of concern in the present study. Although it is plausible that the characteristics of one's occupation influence psychological well-being, it is also possible that persons with low levels of well-being do not want or cannot attain jobs high in complexity, control, and income. For example, someone with low self-confidence may not be considered capable of handling a job with a high degree of control required, or a person with high vulnerability may try to preserve a fragile state of well-being by seeking work that does not make heavy demands on personal resources. The issue of causality cannot be addressed with the data in this study. However, other researchers using longitudinal data have found that changes in occupational conditions can increase productivity and job satisfaction (Katzell & Yankelovich, 1975), reduce psychosomatic complaints (Karasek, 1979; Kohn & Schooler, 1978), and lead to improvements in mental health (Wall & Clegg, 1981); other results contradict the reverse causal hypothesis (Frese, 1985).

In future research the relation between occupational characteristics and well-being might be further clarified by examining the effects of moderator variables in addition to sex. Moderators such as higher order need strength have been examined for their impact on the job characteristics–job satisfaction relation (Hackman & Lawler, 1971; Sims & Szilagyi, 1976; Stone, Mowday, & Porter, 1977) and might also influence the association between occupational characteristics and psychological well-being. For example, in one study, employees with either too little or too much perceived complexity had higher depression than did employees with just the desired amount of complexity in their jobs (Caplan et al., 1975).

Finally, this article began with the observation that a relation between employment status and well-being has been frequently noted. In this study, characteristics of work among those employed for pay were examined. An extension of this work for future consideration would be to test whether the occupational characteristics of housework (and perhaps volunteer work) bear a similar relation to the psychological well-being of full-time homemakers. Stellman (1977), for example, has argued that work in the home "shares most of the worst characteristics of dissatisfying paid employment" (p. 79), including monotony, repetitiveness, and low social and monetary rewards. Perhaps

such differences in occupational characteristics help explain the persistent findings that full-time homemakers have generally lower psychological well-being than do employed men and women.

In conclusion, this study revealed that occupational characteristics are related to the psychological well-being of employed women and men. The results of this study suggest that job enrichment (increasing variety) and enlargement (increasing control) have the potential to influence not only productivity and job satisfaction, but the larger psychological life and well-being of employees as well.

References

- Adelmann, P. K. (1987). *Employment and psychological well-being in women: A meta-analysis*. Unpublished manuscript, University of Michigan, Ann Arbor.
- Bryant, F. B., & Veroff, J. (1984). Dimensions of subjective mental health in American men and women. *Journal of Health and Social Behavior*, 25, 116-135.
- Caldwell, D. F., & O'Reilly, Charles A. III. (1982). Task perceptions and job satisfaction: A question of causality. *Journal of Applied Psychology*, 67, 361-369.
- Campbell, A. (1980). Changes in psychological well-being during the 1970's of homemakers and employed wives. In D. G. McGuigan (Ed.), *Women's lives: New theory, research and policy* (pp. 291-301). Ann Arbor: University of Michigan, Center for Continuing Education of Women.
- Caplan, R. D., Cobb, S., French, J. R. P., Jr., Harrison, R. D., & Pinneau, S. R., Jr. (1975). *Job demands and worker health: Main effects and occupational differences*. Washington, DC: U.S. Government Printing Office.
- Cobb, S., & Kasl, S. (1977). *Termination: The consequences of job loss* (DHEW rep.). Washington, DC: U.S. Government Printing Office.
- Eichar, D. M., & Thompson, J. L. P. (1986). Alienation, occupational self-direction, and worker consciousness. *Work and Occupations*, 13, 47-65.
- Frese, M. (1985). Stress at work and psychosomatic complaints: A causal interpretation. *Journal of Applied Psychology*, 70, 314-328.
- Gardell, B. (1971). Technology, alienation, and mental health in the modern industrial environment. In L. Levi (Ed.), *Society, stress and disease* (Vol. 1, pp. 148-180). London: Oxford Press.
- Hackman, J. R., & Lawler, E. E. III. (1971). Employee reactions to job characteristics. *Journal of Applied Psychology*, 55, 259-286.
- Hackman, J. R., Pearce, J. L., & Wolfe, J. C. (1978). Effects of changes in job characteristics on work attitudes and behavior: A naturally occurring quasi-experiment. *Organizational Behavior and Human Performance*, 21, 289-304.
- Haw, M. (1982). Women, work and stress: A review and agenda for the future. *Journal of Health and Social Behavior*, 23, 132-144.
- Kahn, R. L. (1972). The meaning of work: Interpretation and proposals for measurement. In D. Campbell & R. Converse (Eds.), *The human meanings of social change* (pp. 159-203). New York: Russell Sage Foundation.
- Kahn, R. L. (1981). *Work and health*. New York: Wiley.
- Karasek, R. A., Jr. (1979). Job demands, decision latitude, and mental strain: Implications for job redesign. *Administrative Sciences Quarterly*, 24, 285-308.
- Katzell, R. A., & Yankelovich, D. (1975). *Work, productivity, and job satisfaction*. New York: Psychological Corporation.
- Keon, T. L., & McDonald, B. (1982). Job satisfaction and life satisfaction: An empirical evaluation of their interrelationship. *Human Relations*, 35, 167-180.
- Kish, L., & Hess, I. (1965). *The Survey Research Center's national sample of dwellings*. Ann Arbor: University of Michigan, Institute for Social Research.
- Kohn, M. L. (1969). *Class and conformity: A study in values*. Homewood, IL: Dorsey Press.
- Kohn, M. L., & Schooler, C. (1973). Occupational experience and psychological functioning: An assessment of reciprocal effects. *American Sociological Review*, 38, 97-118.
- Kohn, M. L., & Schooler, C. (1978). The reciprocal effects of the substantive complexity of work and intellectual flexibility: A longitudinal assessment. *American Journal of Sociology*, 84, 24-52.
- Kornhauser, A. W. (1965). *Mental health of the industrial worker*. New York: Wiley.
- Lawler, E. E. III. (1971). *Pay and organizational effectiveness: A psychological view*. New York: McGraw-Hill.
- Martin, J. K., & Hanson, S. L. (1986). Sex, family wage-earning status, and satisfaction with work. *Work and Occupations*, 12, 91-109.
- Miller, J., Schooler, C., Kohn, M. L., & Miller, K. A. (1979). Women and work: The psychological effects of occupational conditions. *American Journal of Sociology*, 85, 66-94.
- Near, J. P., Rice, R. W., & Hunt, R. G. (1978). Work and extra-work correlates of life and job satisfaction. *Academy of Management Journal*, 21, 248-264.
- Near, J. P., Smith, C. A., Rice, R. W., & Hunt, R. G. (1983). Job satisfaction and nonwork satisfaction as components of life satisfaction. *Journal of Applied Social Psychology*, 13, 126-144.
- O'Reilly, C. A. III, & Caldwell, D. F. (1979). Informational influence as a determinant of task characteristics and job satisfaction. *Journal of Applied Psychology*, 64, 157-165.
- Porter, L. W., & Lawler, E. E. III. (1965). Properties of organization structure in relation to job attitudes and job behavior. *Psychological Bulletin*, 64, 23-51.
- Rice, R., Near, J., & Hunt, R. (1979). Unique variance in job and life satisfaction associated with work-related and extra-workplace variables. *Human Relations*, 32, 605-623.
- Ritti, R. (1971). Job enrichment and skill utilization in engineering organizations. In J. Maher (Ed.), *New perspectives on job enrichment* (pp. 131-156). New York: Van Nostrand Reinhold.
- Sims, H. P., & Szilagyi, A. D. (1976). Job characteristic relationships: Individual and structural moderators. *Academy of Management Journal*, 19, 195-212.
- Stellman, J. M. (1977). *Women's work, women's health*. New York: Pantheon Books.
- Stone, E. F., Mowday, R. T., & Porter, L. W. (1977). Higher order need strengths as moderators of the job scope-job satisfaction relationship. *Journal of Applied Psychology*, 62, 466-471.
- Temme, L. V. (1975). *Occupation: Meanings and measures*. Washington, DC: Bureau of Social Science Research.
- Turner, A. N., & Lawrence, P. R. (1965). *Industrial jobs and the worker*. Boston: Harvard Graduate School of Business Administration.
- U.S. Department of Labor. (1965). *Dictionary of Occupational Titles* (3rd ed., 2 vols.). Washington, DC: U.S. Government Printing Office.
- Veroff, J., Douvan, E., & Kulka, R. (1981). *The inner American: A self-portrait from 1957 to 1976*. New York: Basic Books.
- Vroom, V. H. (1964). *Work and motivation*. New York: Wiley.
- Wall, T. P., & Clegg, C. W. (1981). A longitudinal study of group work design. *Journal of Occupational Behavior*, 2, 32-43.
- Warr, P., & Parry, G. (1982). Paid employment and women's psychological well-being. *Psychological Bulletin*, 91, 3, 498-516.
- Weaver, C. N. (1977). Relationships among pay, race, sex, occupational prestige, supervision, work autonomy, and job satisfaction in a national sample. *Personnel Psychology*, 30, 437-445.
- Wright, J. D. (1978). Are working women really more satisfied? Evidence from several national surveys. *Journal of Marriage and the Family*, 40, 301-313.

Appendix A

Definitions of Occupational Characteristics

Temperaments: Different Types of Occupational Situations to Which Workers Must Adjust

Variety: situations involving a variety of duties often characterized by frequent change.

Repetition: situations involving repetitive or short-cycle operations carried out according to set procedures or sequences.

Instruction by others: situations involving doing things only under specific instruction, allowing little or no room for independent action or judgment in working out job problems.

Control of self or others: situations involving the direction, control, and planning of an entire activity or the activities of others.

Complexity of Relationships to Data, People, and Things

Data: information, knowledge, and conceptions related to data, people, or things, obtained by observation, investigation, interpretation, visualization, mental creation; incapable of being touched; written data take the form of numbers, words, symbols; other data are ideas, concepts, oral verbalization.

People: human beings; also animals dealt with on an individual basis as if they were human.

Things: inanimate objects as distinguished from human beings; substances or materials; machines, tools, equipment; products. A thing is tangible and has shape, form, and other physical characteristics.

Data	People	Things
Synthesizing	Mentoring	Setting up
Coordinating	Negotiating	Precision working
Analyzing	Instructing	Operating—controlling
Compiling	Supervising	Driving—operating
Computing	Diverting	Manipulating
Copying	Persuading	Tending
Comparing	Speaking—signaling	Feeding—offbearing
No significant relationship	Serving	Handling
	No significant relationship	No significant relationship

Appendix B

Factor Loadings for Men and Women From Confirmatory Factor Analysis Using Men’s Model of Psychological Well-Being^a

Variable ^b	Unhappiness ^c		Lack of self-confidence ^c		Vulnerability ^c	
	Men	Women	Men	Women	Men	Women
General unhappiness	.81	.81	.00	.00	.00	.00
Past unhappiness	.40	.38	.00	.00	.00	.00
Happiest in past	.38	.35	.00	.00	.00	.00
Low future morale	.72	.67	.00	.00	.00	.00
General dissatisfaction	.42	.44	.00	.00	.00	.00
Zung depression	.00	.00	.73	.73	.00	.00
Low self-esteem	.00	.00	.61	.58	.00	.00
Anomie	.00	.00	.42	.46	.00	.00
Perceived control:problems	.00	.00	.36	.52	.00	.00
Perceived control:outcomes	.00	.00	.30	.33	.00	.00
Low self-acceptance	.00	.00	.23	.24	.00	.00
Frequency of bad things	.00	.00	.00	.00	.47	.48
Frequency of feeling overwhelmed	.00	.00	.00	.00	.68	.68
Nervous breakdown	.00	.00	.00	.00	.51	.60

^a Partial model shown here (3 of 6 factors); full model appears in Bryant and Veroff, 1984.
^b For exact wording of items and indices, see Bryant and Veroff, 1984.
^c Factors were reversed for use in present study.

Received October 17, 1985
Revision received February 17, 1987
Accepted January 26, 1987 ■

Understanding, Prediction, and Control as Moderators of the Relationships Between Perceived Stress, Satisfaction, and Psychological Well-Being

Lois E. Tetrick
Wayne State University

James M. LaRocco
Research Department, Naval School of Health Sciences
Bethesda, Maryland

This study provides a preliminary test of a model proposed by Sutton and Kahn (1986). In the model, the ability to understand, predict, and control events in the work environment can reduce the potential adverse effects generally associated with certain work conditions. Using a sample of physicians, dentists, and nurses ($N = 206$) from a large naval medical hospital, the present study examined the moderating effects of understandable, predictable, and controllable work situations on the relationship between perceived role stress, satisfaction, and psychological well-being. Understanding and control were found to have moderating effects on the relationship between perceived stress and satisfaction. Understanding, prediction, and control were found to have direct relationships with perceived stress, but only control had a significant direct relationship with satisfaction. None of these variables were found to have significant direct relationships with psychological well-being.

The general work stress health model (e.g., House, 1981; Katz & Kahn, 1978) postulates that objective work conditions can lead to perceptions of stress. Perceived stress, in turn, leads to job-related strains such as dissatisfaction, boredom, and turnover, and to individual strains such as anxiety, depression, and physical illness. In addition, the stress health model hypothesizes that internal characteristics (i.e., personal characteristics) and external conditions (i.e., situational characteristics) not only have direct effects, but also have interactive or moderating effects.

Research involving the stress health model has taken three general forms: (a) a demonstration that certain job conditions lead to adverse outcomes (e.g., role ambiguity and role conflict lead to job dissatisfaction); (b) the demonstration of the direct effects on stress and strain of factors external to the workplace (e.g., social support lessens role conflict and depression) or internal to the individual (e.g., Type-A behavior pattern increases role conflict and anxiety levels); and (c) the demonstration of the moderating effects of these internal and external factors

(e.g., social support reduces the relationship between role conflict and depression). The first two lines of research have established a relatively consistent link between perceived role stress, satisfaction, and psychological well-being. However, attempts to confirm the moderating effects of such variables as social support have met with only limited success (Beehr & Bhagat, 1985; Cohen & Wills, 1985). Cohen and Wills (1985) suggested that this limited success may be attributed, at least in part, to the lack of relevancy to the work environment of some of the measures of social support.

Recently, Sutton and Kahn (1986) proposed three work-relevant antidotes against the stress-strain effect in organizational life: understanding of events, predictability of events, and control over events. To the extent that an organizational member can (a) understand how and why events happen, (b) predict the frequency, timing, and duration of events in the work environment, and (c) control the outcomes desired by effectively influencing the events, things, or others in the work environment, less strain will result from the stressors existing in the work environment.

The purpose of this research was to test the potential moderating and main effects of understanding, prediction, and control in the stress-strain relationship. Incorporating Sutton and Kahn's (1986) proposed antidotes into the general work health model, this study hypothesized that (a) understanding, (b) prediction, and (c) control moderate the relations between (1) perceived role stress and job satisfaction (Hypotheses 1a, 1b, and 1c), between (2) perceived role stress and psychological well-being (Hypotheses 2a, 2b, and 2c), and between (3) job satisfaction and psychological well-being (Hypotheses 3a, 3b, and 3c). These hypothesized moderating effects are reflected in Figure 1 by the correspondingly labeled arrows.

It also was hypothesized that (a) understanding, (b) prediction, and (c) control are directly related to all three of the central

This article was supported by the Naval Medical Research and Development Command, Department of the Navy, under Research Work Unit MR 001.RP-018004. The views presented in this article are those of the authors only and do not necessarily represent the official view, policies, or endorsements of the U.S. Navy or any other government agency.

The authors would like to thank the staff of the Medical Psychology Department, Uniformed Services University of the Health Sciences, for their help on this project. The assistance of Laurie Davidson, who was responsible for data collection, and of Andy Baum, who advised on research design, was particularly valuable.

Correspondence concerning this article should be addressed to Lois Tetrick, Department of Psychology, Wayne State University, 71 West Warren, Detroit, Michigan 48202.

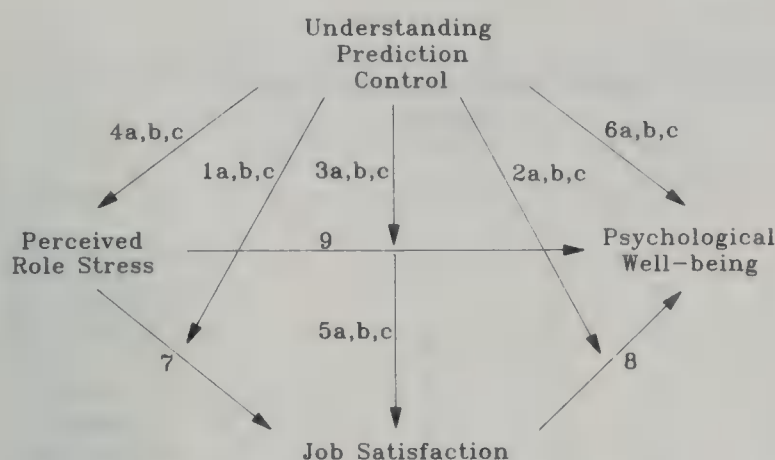


Figure 1. Schematic representation of hypothesized relationships among understanding, prediction, control, perceived role stress, satisfaction, and psychological well-being. (Because the same relationships are hypothesized for understanding, prediction, and control, only one arrow for each set of hypotheses is drawn to maintain clarity and simplicity.)

constructs: (4) perceived role stress (Hypotheses 4a, 4b, and 4c), (5) job satisfaction (Hypotheses 5a, 5b, and 5c), and (6) psychological well-being (Hypotheses 6a, 6b, and 6c). These hypotheses are reflected by the correspondingly labeled arrows in Figure 1. Finally, it was hypothesized that (7) perceived role stress influences job satisfaction (Hypotheses 7, arrow 7), which in turn affects (8) psychological well-being (Hypotheses 8, arrow 8), and that (9) perceived role stress influences psychological well-being directly (Hypotheses 9, arrow 9).

Method

Sample

The data were collected at a large naval hospital in the northeastern United States. Participants were physicians, nurses, and dentists ($N = 225$). On the average, the respondents were in their thirties, had been in their current job approximately 16 months, and were midlevel officers. Most of them were primarily in clinical work as opposed to administration. The physicians and dentists were almost all men, whereas the nurses were almost all women. Approximately half of the nurses were married, and most of the physicians and dentists were married.

Data Collection

Participants were recruited through an announcement at staff meetings and by notes placed in a newssheet published daily at the hospital. Prospective participants assembled at the end of their work day (4–6 p.m.) in a designated room. After being briefed on the study, they reviewed and signed a consent form if they agreed to participate in the research. The participants then received the questionnaire and were instructed to return it the next day. Of the 225 returned questionnaires, 206 contained no missing data and were submitted to analysis. Because this study was part of a larger research project, it was necessary to use abbreviated versions of existing measures.

Measures

Perceived role stress is often operationalized as role ambiguity and role conflict. This study included four items from Caplan, Cobb,

French, Harrison, and Pinneau (1975) to measure role ambiguity and four items from House and Rizzo (1972) for role conflict. The response scale for these items was from 1 = *very little extent* to 7 = *a great extent*.

Job satisfaction was measured using items from Hackman and Oldham (1980). These items were scaled on a seven-point scale ranging from *strongly disagree* to *strongly agree*. Two variables were constructed from these items: a global measure of job satisfaction and a job-facet satisfaction measure, which was a summation of six scales measuring satisfaction with security, pay, growth, co-workers, supervisors, and hours.

Psychological well-being was measured by two scales: anxiety and depression, based on items adapted primarily from Caplan et al. (1975). These items required the respondent to indicate on a five-point scale how frequently they experienced certain physical and psychological responses to work conditions.

Understanding of events, predictability of events, and control over one's work environment (self-determination) were measured by scales developed for this study based on Sutton and Kahn's (1986) definitions. These items are presented in the Appendix. The response scale for the items was from 1 = *very little* to 7 = *a great extent*.

Analysis

Because of the adverse effect of measurement error in testing for interaction effects (Kenny & Judd, 1984), a latent variable model was developed. The measurement model was represented by three latent variables: perceived role stress, job satisfaction, and psychological well-being. The indicators of perceived role stress were role ambiguity and role conflict, the indicators of job satisfaction were global job satisfaction and job-facet satisfaction, and the indicators of psychological well-being were anxiety and depression. Construction of this measurement model does not imply that role ambiguity and role conflict are equivalent but simply implies that there is a common underlying construct—perceived role stress—that is consistent with McGrath's (1976) definition of stress. Similarly, for the other latent variables, no assumption or constraint was imposed that the indicators were equivalent. Understanding, prediction, and control were treated as manifest variables because multiple indicators were not available. Confirmatory factor analysis was used to assess the goodness of fit of the measurement model.

After the adequacy of the measurement model was assessed, one additional confirmatory factor analysis including the additional moderating effects was performed. As described by Kenny and Judd (1984), manifest indicators of each hypothesized latent interactive effect were computed by obtaining the cross-products of the respective manifest indicators for the two latent variables involved in a particular interactive effect. The covariance matrix of all latent variables from this confirmatory factor analysis was used in all subsequent analyses, using LISREL VI.

Creation of the cross-product terms results in a violation of multivariate normality for maximum likelihood estimation procedures (Kenny & Judd, 1984). Boomsma (1983) suggested that small to moderate violations of normality (skewnesses of the observed variables less than 1.0) may not affect parameter estimates, standard errors for testing the statistical significance of parameters, or the chi-square goodness-of-fit statistic. However, only one or two very extreme skewnesses can lead to non-robust results. Therefore, as suggested by Boomsma (1983), Dillon and Goldstein (1984), and Jöreskog and Sörbom, (1986), chi-square values, and the associated probability level, for a given model should not be used to assess the goodness of fit of a population model. Instead, one should use the goodness-of-fit index (GFI), adjusted goodness-of-fit index (AGFI), and, if maximum likelihood or generalized least squares estimation procedures are used, the relative change in chi-square to degrees of freedom of comparative models (Dillon & Goldstein, 1984; Jöreskog & Sörbom, 1986). One can also use a distribution-free estimation

Table 1
Descriptive Statistics and Estimates of Internal Consistency

Variable	N Items	M	SD	Correlations								
				1	2	3	4	5	6	7	8	9
1. Role ambiguity	4	2.88	1.01	.70								
2. Role conflict	3	3.67	1.12	.52	.61							
3. Job satisfaction	3	5.06	1.40	-.44	-.50	.81						
4. Facet satisfaction	15	4.19	0.87	-.46	-.41	.66	.85					
5. Anxiety	5	2.44	0.65	.33	.36	-.41	-.36	.79				
6. Depression	7	2.27	0.58	.34	.35	-.42	-.40	.79	.76			
7. Prediction	3	3.88	0.96	-.27	-.52	.37	.30	-.24	-.28	.66		
8. Understanding	3	4.69	1.06	-.48	-.29	.32	.43	-.22	-.24	.15	.75	
9. Control	6	4.21	1.15	-.46	-.39	.38	.48	-.24	-.26	.21	.50	.83

Note. Numbers in boldface represent coefficient alphas.

procedure. LISREL VI provides unweighted least squares estimation as the only distribution-free estimation procedure; only the GFI and AFGI are provided for this procedure.

Bentler and Bonett's (1980) procedure for testing the statistical significance of the hypothesized relations was used. This procedure involves estimating the fit of a series of sequential, nested models in which each model is progressively more restricted (one less parameter to be estimated). The difference in chi-square between two sequential models is analogous to the change in R^2 in moderated hierarchical regression; however, the order is the reverse of moderated regression. One starts with a model in which the cross-product terms are included and then reestimates the model with the parameter for a given cross-product term fixed to zero. Then one compares the relative fit of the two models. If the chi-square difference obtained from subtracting the chi-square value for the former model from the chi-square value for the more restricted model is statistically significant, the hypothesized effect is supported. This sequence of testing nested models was continued until all moderating effects and then direct relationships were tested.

Results

The number of items and the estimates of internal consistency based on coefficient alpha are shown in Table 1. The means and standard deviations for the nine variables included in this study and the zero-order correlation coefficients among the variables also are presented in Table 1. None of the observed variables' skewnesses were greater than 1.0, ranging from .02 to .47 in absolute value. The range of skewnesses for the cross-product indicators was from .17 to .75. The lack of large skewnesses indicates that maximum likelihood estimation procedures comparing relative fit between models are not inappropriate (Baron & Kenny, 1986; Boomsma, 1983).

Examination of the correlation matrix indicates that role conflict and role ambiguity were correlated ($r = .52$). Both of these variables were correlated in similar directions with all the remaining variables. However, role conflict was more negatively correlated with prediction ($r = -.52$) than was role ambiguity ($r = -.29$), $t(203) = -4.15$, $p < .05$. Role ambiguity was more negatively correlated with understanding ($r = -.48$) than was role conflict ($r = -.29$), $t(203) = -3.39$, $p < .05$. Facet satisfaction and job satisfaction were strongly correlated with each other ($r = .66$) as were anxiety and depression ($r = .79$). Understanding and control were correlated ($r = .50$), whereas predic-

tion was only weakly correlated with understanding or control ($r = .15$ and $.21$, respectively).

The measurement model was tested by confirmatory factor analysis using LISREL VI. The resulting lambda parameter estimates and the correlations among the reduced variables are shown in Table 2. The chi-square value of 35.44 with 21 *dfs* ($p = .025$) indicates that the imposed dimensionality among the observed variables did not account for all of the covariance. However, the ratio of chi-square to degrees of freedom (less than 2:1), the root mean residual (.053), and the GFI (.962) indicate that the hypothesized measurement model adequately accounted for the covariance among the observed variables (Bentler & Bonett, 1980; Jöreskog & Sörbom, 1986).

Table 3 reflects a summary of the test of sequential, nested models addressing the hypothesized effects for understanding, prediction, and control, each sequential model being more restricted than the previous model. The coefficients of determination for each of the endogenous variables (perceived role stress, satisfaction, and psychological well-being) and the total system of equations are shown. The chi-square, GFI, and AGFI also are shown for each more restricted model.

Examination of the results of the tests of the sequential, nested models for the moderating effects of understanding, prediction, and control indicated no significant effect on the relationship between satisfaction and psychological well-being (Hypotheses 3a, 3b, and 3c) or on the relationship between perceived role stress and psychological well-being (Hypotheses 2a, 2b, and 2c). Progressively restricting the model by fixing the parameters for the cross-product terms in the psychological well-being equation to zero did not result in a statistically significant chi-square difference. After deleting these six parameters and gaining 6 *dfs*, the chi-square value had increased only 4.90 from 134.38 to 139.28.

Similarly, eliminating the hypothesized moderating effect of prediction on the relationship between perceived role stress and satisfaction (Hypothesis 1a) did not result in a significant chi-square difference (1.44, with 1 *df*). However, further restricting the model to eliminate the moderating effects of understanding and control on the relationship between perceived role stress and satisfaction yielded significant chi-square differences (9.22 and 7.90, respectively).

Table 2
Parameter Estimates From Measurement Model for Nine Original and Six Reduced Variables

Variable	Perceived role stress	Satisfaction	Psychological well-being	Prediction	Understanding	Control
Lambda parameter estimates						
Role ambiguity	.74					
Role conflict	.76					
Job satisfaction		.82				
Facet satisfaction		.82				
Anxiety			.88			
Depression			.92			
Prediction				—		
Understanding					—	
Control						—
Correlations among reduced variables (phi matrix estimates)						
Perceived role stress	—					
Satisfaction	-.73	—				
Psychological well-being (reflected)	.51	-.55	—			
Prediction	-.58	.46	-.32	—		
Understanding	-.57	.48	-.29	.28	—	
Control	-.59	.54	-.30	.28	.50	—

Note. χ^2 (21, $N = 206$) = 35.44, $p = 0.025$. Goodness-of-fit index = 0.962. Root mean residual = 0.053.

Retaining only the two moderating effects found to be significant, the hypothesized direct effects of understanding, prediction, and control on psychological well-being (Hypotheses 6a, 6b, and 6c) were tested. None of the chi-square difference tests were significant, indicating that understanding, prediction, and control did not have direct relationships with psychological well-being.

Hypothesis 5 stated that understanding, prediction, and control were directly related to satisfaction. The chi-square difference test indicated that there was no direct relationship between prediction and satisfaction (change in chi-square was 2.70, with 1 *df*; $p > .05$). Similarly, adding the constraint that the parameter linking understanding to satisfaction was zero did not result in a significant chi-square difference (change in chi-square was 2.47; $p > .05$). However, the chi-square difference eliminating the linkage between control and satisfaction was significant (change in chi-square was 6.92; $p < .05$), supporting this hypothesized relationship. All other direct hypothesized relationships were supported.

The results of this sequential test of nested models only partially supports the hypothesized relationships shown in Figure 1. Perceived role stress had a direct relationship with satisfaction and psychological well-being. Satisfaction had a direct relationship with psychological well-being. Understanding, prediction, and control all had direct relationships with perceived role stress and no direct relationships with psychological well-being. Only control had a direct relationship with satisfaction and only understanding and control were found to moderate the relationship between perceived role stress and satisfaction. None of the other hypothesized moderating effects were supported.¹

Discussion

This study provides partial support for Sutton and Kahn's (1986) proposal that understanding, prediction, and control can

be "antidotes" to adverse work conditions. Understanding and control were found to moderate the negative relationships between perceived role stress and satisfaction. Support was not found for a moderating effect of understanding, prediction, and control on the relationship between perceived role stress and psychological well-being or the relationship between satisfaction and psychological well-being.

Understanding and control may be most useful in moderating the effects of organizational conditions on job-related attitudes and strains and least effective when health outcomes are of concern, although understanding, prediction, and control all had indirect effects on psychological well-being through their effects on perceived role stress. This finding, if substantiated in future research, is intriguing because research on the moderating effect of social support tended to find just the opposite. That is, social support generally has been found to moderate the effects of organizational conditions on psychological and physical health, but generally it has not been found to moderate these effects on job-related attitudes and strains (LaRocco, House, & French, 1980).

One explanation for the moderating effect of understanding and control found in this study and those found for social support in other studies may involve the principle of relevancy (Cohen & Wills, 1985; French, Caplan, & Harrison, 1982). The principle of relevancy holds that the strongest relationship between an independent and dependent variable occurs when both are measured on commensurate dimensions. By extrapo-

¹ The results of the analyses were not substantially different for solutions based on generalized least squares or unweighted least squares estimation procedures. For generalized least squares, this model resulted in the largest adjusted goodness-of-fit index (AGFI; .803), with a chi-square of 40.44. Unweighted least squares estimation procedures also resulted in this model having the largest AGFI (.985).

Table 3
Test of Moderating and Direct Relation Based on a Comparison of Sequential, Nested Models

Sequential model	Coefficient of determination				Goodness-of-fit indices			
	Role stress	Satisfaction	Psychological well-being	Total	df	χ^2	GFI	AGFI
Sequential tests of moderating effects of understanding, prediction, and control								
Full model	.635	.647	.352	.674	9	134.38	.921	.316
Moderating effect on relation between								
Role stress and well-being	.635	.647	.352	.684	12	135.36	.920	.482
Satisfaction and well-being	.635	.647	.339	.678	15	139.28	.918	.572
Role stress and satisfaction								
Prediction \times Role Stress	.635	.646	.339	.676	16	140.72	.917	.618
Understanding \times Role Stress	.635	.613	.330	.667	17	149.94*	.917	.618
Control \times Role Stress	.635	.620	.330	.666	17	148.62*	.916	.614
Sequential tests of direct effects of understanding, prediction, and control								
Modified model ^a	.635	.646	.339	.676	4	34.29	.963	.663
Direct relation with								
Psychological well-being	.635	.646	.332	.673	7	36.02		
Satisfaction							.960	.797
Prediction	.635	.646	.332	.673	8	36.02	.960	.797
Understanding	.635	.642	.332	.671	9	38.49	.958	.831
Control	.635	.642	.335	.664	10	45.41*	.951	.824
Role stress								
Prediction	.470	.643	.332	.521	10	115.07*	.900	.640
Understanding	.546	.645	.333	.589	10	83.34*	.921	.717
Control	.595	.628	.325	.638	10	60.03*	.939	.779
Sequential tests of direct relation among perceived role stress, satisfaction, and psychological well-being								
Role stress and well-being	.635	.352	.305	.671	10	46.99*	.947	.808
Satisfaction and well-being	.635	.642	.275	.671	10	53.11*	.943	.796
Role stress and satisfaction	.635	.642	.271	.729	10	151.24*	.880	.568

Note. The sequence of models represents progressively more restricted models (i.e., there are fewer parameters free to be estimated). Significant increases in χ^2 relative to the change in *df* indicate that the parameters that have been fixed should be left in the model rather than being fixed to zero. GFI = goodness-of-fit index; AGFI = adjusted goodness-of-fit index.

^a Modified model is based on the covariance matrix containing only those cross-product terms that were significant.

* Change in χ^2 is significant, $p < .05$.

lation, one might hypothesize that moderator variables are likely to have more and greater effects to the extent that their conceptual dimensions are relevant to the independent and dependent variables whose relationship they are said to moderate. For example, social support, which is almost universally measured as socioemotional in nature, more often moderates relationships involving socioemotional outcomes, such as anxiety, depression, and somatic complaints. Perceived understanding and control, being perceived job characteristics and more relevant to events at work, appear to moderate relationships involving other job characteristics and attitudes such as job satisfaction.

Another interesting aspect of this study was the differential relationships between both understanding and prediction, on the one hand, and both role ambiguity and role conflict, on the other. Jackson and Schuler (1985) argued, based on their meta-analysis, that role conflict and role ambiguity, although generally correlated with each other, are differentially related to other variables. Support for their argument is evident in the zero-order correlations in this study relative to understanding and prediction but not with the satisfaction or psychological well-being

variables. Predictability of events in the work environment was more negatively correlated with role conflict than role ambiguity, whereas understanding of events in the work environment was more negatively related to role ambiguity than role conflict. Control was correlated with both role conflict and role ambiguity. One possible explanation of these relationships is that perceived predictability may lead to an illusion of control (Langer, 1983), at least within a larger time frame. This could allow an individual to avoid simultaneously competing demands and thus reduce perceived role conflict. Understanding, on the other hand, may reduce ambiguity because one knows why events are occurring—but knowing why things occur does not necessarily provide a means to avoid or spread out demands on one's resources.

This pattern of relationships also suggests that the psychologically important aspects of one's role in the work environment may be understanding and control. Deci and Ryan (1985) have suggested that self-determination (control over the work environment) and perceived self-competence (similar to understanding) are the central motivational variables in the work environment. In light of the results of the study reported here, as

well as other investigations of organizational stress (Lazarus & Folkman, 1984; McGrath, 1976), an integration of the theories and empirical results on stress and motivation in the workplace may help clarify the inconsistent findings in both of these research areas.

Summary

This study found support for Sutton and Kahn's (1986) proposed moderating effects of understanding and control on the relationship between role stress and satisfaction. Prediction was not found to moderate this relationship. Understanding, prediction, and control did not moderate the relationship between perceived role stress and psychological well-being or between satisfaction and well-being. Understanding, prediction, and control were found to be directly related to perceived role stress. Control also was directly related to satisfaction. Therefore, it would appear that understanding of events, predictability of events, and control over outcomes in the work environment may serve as antidotes to occupational stress.

References

- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173-1182.
- Beehr, T. A., & Bhagat, R. S. (1985). *Human stress and cognition in organizations: An integrated perspective*. New York: Wiley.
- Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588-606.
- Boomsma, A. (1983). *On the robustness of LISREL (maximum likelihood estimation) against small sample size and non-normality*. Unpublished doctoral dissertation. Rijksuniversiteit, Psychologische Institution, Groningen, Netherlands.
- Caplan, R. D., Cobb, S., French, J. R. P., Jr., Harrison, R. V., & Pinneau, S. R., Jr. (1975). *Job demands and worker health* (HEW Publication No. NIOSH-75-160). Washington, DC: National Institute of Occupational Safety and Health.
- Cohen, S., & Wills, T. A. (1985). Stress, social support, and the buffering hypothesis. *Psychological Bulletin*, 98, 310-357.
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.
- Dillon, W. R., & Goldstein, M. (1984). *Multivariate analysis: Methods and applications*. New York: Wiley.
- French, J. R. P., Caplan, R. D., & Harrison, R. V. (1982). *The mechanisms of job stress and strain*. New York: Wiley.
- Hackman, J. R., & Oldham, G. (1980). *Work redesign*. Reading, MA: Addison-Wesley.
- House, J. S. (1981). *Work stress and social support*. Reading, MA: Addison-Wesley.
- House, R. J., & Rizzo, J. R. (1972). Role conflict and ambiguity as critical variables in a model of organizational behavior. *Organizational Behavior and Human Performance*, 7, 467-505.
- Jackson, S. E., & Schuler, R. S. (1985). A meta-analysis and conceptual critique of research on role ambiguity and role conflict in work settings. *Organizational Behavior and Human Decision Processes*, 36, 16-78.
- Jöreskog, K. G., & Sörbom, D. (1986). *LISREL VI: Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least square methods* (4th ed.). Mooresville, IN: Scientific Software.
- Katz, D., & Kahn, R. L. (1978). *The social psychology of organizations* (2nd ed.). New York: Wiley.
- Kenny, D., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, 96, 201-210.
- Langer, E. J. (1983). *The psychology of control*. Beverly Hills, CA: Sage.
- LaRocco, J. M., House, J. S., & French, J. R. P., Jr. (1980). Social support, occupational stress, and health. *Journal of Health and Social Behavior*, 21, 202-218.
- Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal and coping*. New York: Springer.
- McGrath, J. E. (1976). Stress and behavior in organizations. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 1351-1395). Chicago: Rand McNally.
- Sutton, R., & Kahn, R. L. (1986). Prediction, understanding, and control as antidotes to organizational stress. In J. Lorsch (Ed.), *Handbook of organizational behavior*. Englewood Cliffs, NJ: Prentice-Hall.

Appendix

Items Used to Measure Understanding of Events, Predictability of Events, and Control Over One's Work Environment

Understanding of events

- To what extent do you know why others at work act as they do?
- To what extent do you understand the reason organizational changes occur?
- To what extent do you understand the reasons why job-related decisions were made?

Predictability of events

- To what extent can you predict what job demands will be placed on you each day?
- To what extent do unexpected events occur on your job?
- To what extent are you faced with unexpected decisions concerning your work?

Control over one's work environment

- To what extent do you have influence over the things that affect you on the job?
- To what extent do you have input in deciding what tasks or parts of tasks you will do?
- To what extent do you have the opportunity to take part in making job-related decisions that affect you?
- To what extent can you set your own work deadlines?
- To what extent does your job allow you the opportunity for independent thought and action?
- To what extent do you control the pace and scheduling of your work?

Received July 11, 1986

Revision received April 6, 1987

Accepted April 28, 1987 ■

Pay, Equity, Job Gratifications, and Comparisons in Pay Satisfaction

Leonard Berkowitz
University of Wisconsin

Colin Fraser and F. Peter Treasure
University of Cambridge, Cambridge, England

Susan Cochran
University of Wisconsin

In an investigation of the determinants of pay satisfaction, we held telephone interviews with 248 fully employed men in Dane County, Wisconsin, asking about their income, job satisfaction, and other economic and demographic matters. The social and industrial-organizational psychology literature suggests that pay satisfaction could be influenced by at least four major considerations: the economic benefits received on the job, the extent to which earnings are regarded as fair or deserved, comparisons with other people's pay, and noneconomic job satisfactions. Measures of these possible determinants were established by a factor analysis of 29 items, and the index of pay satisfaction was based on another factor analysis of 8 items. Using these factors and several demographic variables in a multiple regression analysis, we found that three of our four types of psychological determinants made major contributions to predicting pay satisfaction, with the most powerful set of predictors being equity considerations, although material benefits associated with living standards and intrinsic job satisfaction were also major predictors. Social comparisons contributed virtually nothing. Small, significant effects were found for age, occupation, education, and past unemployment. Theoretical implications of the findings are discussed.

Satisfaction with pay, as with the attainment of any valued outcome, is likely to be a function of several different processes. At the simplest level, people could respond fairly directly to the money itself. Earnings permit them to purchase the goods and services they desire, and as a consequence, the greater their income, the stronger should be their satisfaction. In a somewhat more complicated manner, they might also evaluate their pay in terms of a standard regarding these economic benefits. One standard is a sense of equity. Are they getting what they deserve? Another standard involves social comparisons. Is their pay as much as someone else's? In this case, apparently, it is not the absolute value of the earnings that is considered so much as the degree to which this outcome meets the relevant standard. Yet another process involves satisfaction with some other aspect of their job. Positive feelings could generalize from, say, intrinsic job satisfaction to their pay, or a conscious trade-off could be accepted in which one satisfaction substitutes for another.

As we will show, on the basis of theoretical grounds and empirical evidence, it is reasonable to expect that any or all of these four distinguishable processes should influence people's sense of satisfaction with their pay. To the best of our knowledge, however, these four plausible determinants have not been simultaneously examined in prior research. Furthermore, many stud-

ies of pay satisfaction have, for practical reasons, used limited segments of the work force, for example, employees of one large firm or organization (e.g., Finn & Lee, 1972; Goodman, 1974; Ronen, 1986) or one type or level of employee in a variety of firms (e.g., Andrews & Henry, 1963; Dyer & Theriault, 1976). Because of this restricted sampling, however, the generality of the findings is not always clear. The results conceivably might apply only to a particular type of employee or organization. In examining the interplay of the four major types of processes in wage satisfaction, it seemed desirable to us to first establish a general picture by studying a random sample of a labor force before undertaking more specialized studies of limited subsamples. Such a procedure would permit the investigation of the relative importance of the four psychological processes in the context of a fuller set of demographic variations than has typically been available in previous studies. Our aim, then, was to conduct an exploratory study of psychological processes and demographic characteristics as determinants of pay satisfaction in a random sample of full-time male workers in one labor market.

We will now consider the possible psychological mechanisms in more detail. Several lines of evidence, having to do with the absolute value of one's economic benefits, indicate that the greater economic and social pleasures produced by a higher income can at times override other considerations, such as whether these earnings are just or fair. According to a number of experimental tests of equity theory, for example, many persons are willing to depart from principles of what is a just return, in order to obtain as favorable an outcome as possible for themselves (e.g., Leventhal & Anderson, 1970). Their desire for substantial benefits frequently outweighs their desire for an equitable return on their inputs. A number of studies (e.g., Dyer & Theriault, 1976; Ronan & Organt, 1973; Schwab & Wallace, 1974), more directly relevant to pay satisfaction, have reported

This research was supported by a grant from the Vilas Foundation of Madison, Wisconsin, to the first author, together with support from Churchill College, Cambridge, England, to the second author.

We wish to thank David Good, Daniel Katz, Catherine Marsh, and Susan Salkield for their helpful comments and assistance.

Correspondence concerning this article should be addressed to Leonard Berkowitz, Department of Psychology, University of Wisconsin, 1202 West Johnson Street, Madison, Wisconsin 53706.

that people's wage level significantly predicts how satisfied they are with their income, even without any measurement of discrepancies from a supposedly fair level of pay. Higher earnings give them a greater ability to enjoy what they want out of life and are gratifying in themselves. However, it is worth noting, as Motowidlo (1982) has observed, that the correlation between income level and pay satisfaction is typically fairly moderate (with the r s ranging from .13 to .46). Frequently, other considerations, besides the absolute value of one's earnings, also influence attitudes toward satisfaction with pay.

The most frequently discussed of these considerations has to do with standards of justice and notions of what is a deserved level of pay. As equity theorizing has long maintained (Adams, 1965; Lawler, 1971), many individuals have some idea of what is a fair wage for them, and are bothered if their income does not at least meet this minimum standard. Several writers have suggested that equity standards determine what is the lowest level of reward that would be regarded as fair or satisfying (Messé & Watts, 1983; Thibaut & Kelley, 1959). The greater the outcome above this minimum criterion, the more satisfying it would presumably be (Messé & Watts, 1983).

There is a major (and still unresolved) problem here, however. Equity analyses are united in assuming that dissatisfaction with one's pay increases, the greater the wage's discrepancy from the deserved pay level, but they disagree as to how much of this latter standard rests on comparisons with other persons. For Adams (1965), Lawler (1971), and others (e.g., Crosby, 1976; Martin & Murray, 1983; Schwab & Wallace, 1974), the standard used in deciding what is a just return is based to a great extent on social comparisons—at times with a variety of different persons (see Ronen, 1986). Presumably, people judge whether the ratio of their outcomes to their inputs matches the outcome to input ratios of other persons in similar circumstances, although not necessarily in the same organization (Ronen, 1986). Several lines of evidence attest to the role of these social comparisons. Investigations in naturalistic settings have shown, for example, that employees' satisfaction with their economic rewards is influenced to a considerable degree by a comparison with the rewards obtained by others (Goodman, 1974; Goodman & Friedman, 1971; Patchen, 1961). And similarly, in a laboratory experiment, university students' sense of being unfairly paid (when they received less money than they thought they deserved) was exacerbated when similar others were given higher wages (Messé & Watts, 1983).

Nonetheless, several theorists (e.g., Davies, 1962; Gurr, 1970; Jaques, 1961) have proposed that judgments of what is a fair return can also be affected by largely internal standards. In one variation on this theme, both Davies and Gurr maintained that political unrest frequently arises when there is a substantial discrepancy between present outcomes and expectations based on past outcomes, whereas in another version, Jaques (1961) argued that wage earners base their notion of what they deserve largely on characteristics of their job. Following this internal standards position, employees conceivably might believe their job performance or any other input, such as their training and experience, should determine their financial compensation somewhat independently of what others in their organization might be getting (Dyer & Theriault, 1976; Finn & Lee, 1972).

There is also another kind of ambiguity regarding the role of

social comparisons. Because of the way in which equity ideas have been linked with social comparisons, it is often assumed that these comparisons influence wage satisfactions only by affecting judgments of what is a fair pay. This is an unduly restrictive conception. In his pioneering study of the psychology of status, Hyman (1942) noted that people assess their position on most social hierarchies by comparing themselves with certain other groups fairly directly and without considering what is deserved or fair. Similarly, Festinger's (1954) theory of social comparison processes makes no reference to fairness or justice. His formulation holds that substantial discrepancies from similar or attractive others will be bothersome independently of the presumed fairness of the given ranking. Putting these arguments together, wage earners could conceivably compare their income with that of others in their social group without asking whether anyone's pay was just. Independently of what anyone might deserve, they theoretically try to determine whether they are getting more, the same, or less money than their reference group, and presumably would feel bad if they saw they were earning less. What matters most is that they were not "keeping up with the Joneses."

A fourth conceivable influence on pay satisfaction is the satisfaction, or lack of it, derived from other facets of the job, such as intrinsic satisfaction with the content of the work. One possibility is that people engage in conscious trade-offs, sacrificing some possible sources of pleasure for other gratifications. It is well-known, at least to academics, that the main reason we academics are not all highly paid business executives is that we are unwilling to give up the joys of the intellectual life and the creativity and autonomy involved in seeking after truth. As long as we can retain these intrinsic satisfactions, we are happy with the lesser salaries of academia. Other groups of workers, with less exalted sentiments and less choice in the labor market, might consciously reconcile themselves to relatively low pay if their hours are good and their surroundings are pleasant. An alternative to such a cognitive analysis is that there is a generalization of feeling from one aspect of the job to another. Such a generalization phenomenon is well-known, of course, and is involved in many different domains of thought and behavior, including cognitive consistency (Heider, 1958), the development of attitudes (see, e.g., Clore & Byrne, 1974), and the influence of moods on judgments and behavior (Isen, 1984). However this phenomenon is manifested and, whatever it might be termed in any specific instance, in all of these cases the affect aroused by some event spreads to other salient features of the situation. If such an effect influences attitudes toward one's income, the pleasure (or reinforcements) people obtain from some aspects of their lives (such as from the intrinsic nature of their work) might color their opinion of their other life outcomes. Feeling good about their job, they might feel good about their pay.¹

Despite these plausible grounds for expecting that some non-pay aspects of work would affect people's satisfaction with their

¹ This is not to say that the pleasure from the intrinsic aspects of the job must invariably influence pay satisfaction. If the intrinsic gratifications are not very strong or if the employees' desires for high wages and the benefits of high pay are relatively great, the generalization of feelings from job to pay may not occur.

earnings, very few direct empirical studies of the possibility appear to exist (e.g., Weitzel, Harpaz, & Weiner, 1977). A routine assumption has been that wage satisfaction will be one of a number of determinants of more general job satisfaction (Lawler, 1971). The possibility that aspects of job satisfaction can influence pay satisfaction has been largely ignored.

Given all of the possibilities just discussed and the uncertainties regarding the different processes that might be operating, it is clearly still too early to offer a precise theoretical model of the determinants of income satisfaction. We have essentially suggested that four types of factors could affect people's satisfaction with their earnings: (a) the magnitude of the economic benefits they receive (such as the absolute value of their pay and the standard of living they enjoy because of their income), (b) their equity considerations (the discrepancy from the pay they think they deserve), (c) social comparisons (that may or may not be linked to ideas of a just wage), and (d) other satisfactions they obtain from work (such as their satisfaction with the intrinsic nature of their job). However, we cannot say which of these determinants is most important or how they might be interconnected. The present investigation is, therefore, only a preliminary exploration. Using measures designed to tap each of these determinants, we ask whether any or all of them predict income satisfaction. In addition, we assessed the relative contributions of these determinants as compared with the contributions made by a number of standard demographic characteristics of wage earners.

Method

Sample

Following their standard procedure, interviewers at the University of Wisconsin Survey Research Laboratory, in the autumn of 1981, made 1,210 telephone calls through random digit dialing and reached 768 actual phone numbers in the surrounding Dane County. If there was no reply, interviewers called the number again until someone answered. When the connection was with a noncommercial party, interviewers identified themselves and asked to speak to a full-time employed adult male over 18 years of age. (The sample was restricted to men partly because of limited funds and partly because of our decision not to include part-time workers.) There was no adult man at the number called in 165 instances out of the 609 completed noncommercial calls; a man was available, but not working full time, in 130 instances. If a fully employed man was at the number but was not available at the moment, the interviewer called back until that person could be reached. Of the 314 eligible cases, 66 refused to be interviewed, resulting in a total random sample of 248 interviewed respondents (i.e., 79% of those eligible).

The sample was quite heterogeneous, despite the fact that Dane County is the seat of both the state government and a major university, and has a decided young, middle-class composition. Most of the respondents were fairly young, although the sample as a whole had a wide range of ages; almost one third were in their 30s and about 30% were 29 years of age or younger. About 19% of the men were in their 40s and the remaining 18% ranged from 50 to 68 years of age. Of the men interviewed, 8% had not completed high school and 38% were high school graduates only. One third of the men had at least a college or university degree, and about 10% had a professional degree (MD, DDS, JD, or PhD). As another example of the heterogeneity of the sample, 38% regarded themselves as being in either the "working" or "lower" social class and 56% indicated that they were members of the "middle" class. Only 2% did not classify themselves.

There was also a considerable range in the men's reported annual income: from about \$5,400 to more than \$70,000. Of the respondents, 7% either refused to report or said they did not know their income. Of those who revealed their yearly pay, approximately one half indicated that their income was about \$21,000 or less, and about 14% said that they had an annual income of more than \$40,000.

In other respects, however, the sample was relatively homogeneous. Reflecting the racial composition of Dane County, it was virtually all White; only 7 nonwhite respondents were included. Two-thirds of the men were married at the time of the interview, and only about 17% had never been married. There was one other wage earner, besides the respondent, in one half of the households in the sample, and about 39% of the men indicated that they alone had any real income.

Interview

After telling the respondents that the survey had mostly to do with people's views about their jobs and assuring them that their replies would be confidential, the interviewer posed some 92 questions that were mostly about the respondent's perception of and attitudes toward his work and income. The interview schedule had been devised by the first two authors. Most of the questions were fixed alternative items that could be answered readily and quickly.² Because of the structured nature of the great majority of questions, one half of the interviews were completed in 20 min or less and about 90% took one-half hour or less. The remaining 10% of the interviews lasted as long as one hour.

Principal Measures

Incomplete data. To save as many cases as possible for the data analyses, we decided to exclude only those respondents who had not answered more than 10% of the questions. For those items constituting an explicit continuum, missing data were filled in by inserting the mean on the particular measure for the missing cases. However, because every question did not provide such an explicit response continuum, the analyses are based on different samples. The major regression analysis reported below used 239 cases, whereas the factor analysis of the psychological items involved 230 cases.

Data reduction. The interview schedule contained questions on issues over and above pay satisfaction and its four likely psychological precursors. There were, however, 37 questions that appeared to relate to one or other of these five issues. Some data reduction was desirable. First, we considered direct measures of pay satisfaction. The remaining 29 questions were then subjected to a factor analysis to determine whether clusters of items reflecting our four conjectured psychological issues could be obtained. An extracted factor was accepted only if its eigenvalue was equal to or greater than one.

Pay satisfaction. Two items asked about overall pay satisfaction, and six items attempted to explore different aspects of wage satisfaction. All eight were subjected to a principal component analysis without rotation. The analysis extracted one component that accounted for 66.2% of the variance. The loadings of the eight items varied from .86 to .69. Although some investigators (e.g., Weiner, 1980) have used measures of different kinds of pay satisfaction, these results clearly showed that all of the items inquiring into respondents' satisfaction with pay clustered in a single component.

² Standard attitude assessment scales such as the Minnesota Satisfaction Questionnaire were not used in this investigation because we wanted a relatively small number of items that would be suitable for our broad range of respondents employed in a wide variety of occupations. We believed that some of the job and pay aspects assessed by these standard scales could safely be ignored in the present exploratory study.

Table 1
Rotated Factor Loadings of Each Item

Factor and item	Eigenvalue	1	2	3	4	5	6	7	8	Communality
1. Social Comparison Frequency with	7.20									
Others paid same		.91	.01	.03	-.06	-.03	-.00	.03	-.02	.85
Particular groups		.91	.01	-.04	-.05	-.00	.02	-.08	.02	.85
Others paid more		.91	.04	-.09	-.02	-.02	.05	-.04	.02	.83
Others paid less		.91	-.01	.05	-.01	.02	-.04	-.07	-.00	.84
People in other occupations		.89	.02	.00	.04	-.09	.02	-.09	.07	.81
People in other organizations		.88	-.04	.02	.02	.09	-.01	.05	-.04	.78
Certain other individuals		.88	-.02	-.03	.05	-.03	.07	.06	.08	.77
People in same occupation		.84	-.01	-.05	-.05	.09	-.05	.01	-.07	.73
People in same organization		.84	-.00	-.01	.04	-.06	-.07	.14	-.03	.73
2. Intrinsic Job Satisfaction	4.14									
Work interesting		-.01	.87	.01	-.01	-.19	.04	-.04	.08	.76
Can use special abilities		.03	.83	-.05	-.06	.05	-.12	-.07	.02	.76
Satisfaction with job in general		-.07	.73	-.25	.02	.15	-.10	-.07	-.04	.70
Supervisors treat me well		.05	.54	.20	.04	-.05	-.21	.07	.14	.48
Promotion chances good		-.00	.53	.16	.05	.33	-.05	.27	-.22	.57
3. Current Inequity	1.78									
Difference from deserved pay/yearly pay		-.04	.10	.71	-.07	-.34	.17	.30	.15	.72
Yearly pay		.13	.30	-.59	-.08	-.13	.16	.14	.09	.54
Getting pay deserved		-.07	-.02	-.54	.10	.27	.20	.38	.07	.63
Unfair certain others paid same		.08	.08	.51	-.01	.28	.42	-.09	-.08	.56
4. Total Household Income	1.45									
Household income		-.05	.01	-.02	.77	.03	.07	-.18	-.04	.64
Part-time income		.01	-.04	-.02	.75	-.17	-.14	.23	.07	.66
5. Non-Pay Economic Benefits	1.36									
Fringe benefits good		.01	-.09	-.03	-.12	.77	-.07	.14	.12	.65
Job security good		-.09	.17	-.07	.03	.61	-.06	-.04	.17	.52
6. Satisfaction with Work Environment	1.28									
Coworkers pleasant		.08	.13	.05	.06	.06	-.76	.01	-.05	.65
Physical surroundings pleasant		-.05	.11	-.02	-.03	.10	-.74	.02	.09	.65
7. Future Equity	1.27									
Future chance of getting deserved pay		.09	.01	.08	.21	.16	-.01	.75	-.05	.64
Unfair, certain others paid more		.10	.02	.13	.32	.08	.12	-.67	-.06	.64
8. Quality of Life	1.13									
Own standard life versus parents'		-.07	-.02	-.08	-.01	-.02	-.06	.08	.73	.59
Work hours good		.12	-.01	.20	-.02	.18	-.03	-.13	.69	.55
Satisfactory standard of living		-.09	.30	-.21	.11	.11	.11	.06	.48	.52

Factor Analysis of Predictor Items

The factor analysis with oblique rotation of the remaining 29 items (using the OBLIMIN option on SPSS^a) extracted eight factors, which in total accounted for 67.7% of the variance. Table 1 summarizes these factors and lists the loadings of each item, after rotation, on the factors.

Because we had been especially interested in the possible role played by social comparisons, nine items asked how frequently respondents compared their pay with others of different types. All of these questions entered into Factor 1, Frequency of Social Comparisons. Thus, someone who claimed to make comparisons with people in his own organization was also likely to say that he made comparisons with people in other organizations, in the same and different occupations, with those who were paid more and with those paid less, and so on. Although some researchers (e.g., Ronen, 1986) required their respondents to differentiate between comparison groups, in the light of these data it would be unjustified in future research on the selection of comparison others to assume that different types of comparisons are necessarily mutually exclusive.

Factor 2 was composed of five of the eight items dealing with job satisfaction. The items loading highest on this factor had to do with the extent to which the work was interesting and gave the men an opportunity to use their own special abilities, and we therefore regarded the

factor as a measure of the respondents' satisfaction with the intrinsic aspects of their jobs.

Factor 3 involved the degree to which the men viewed their current pay as inequitable. The item with the highest loading was a quantitative index composed of the difference between the pay they believed they deserved and their present (annual) pay divided by their annual wage. Weiner (1980) had found that this measure, which she termed *relative equitable pay*, correlated more highly with her index of pay satisfaction than was the absolute difference between received and deserved pay. However, the respondents' actual yearly pay was also related (moderately negatively) to this current inequity factor, indicating that those who were relatively poorly paid tended to see their wage as unfair. The remaining two items in this factor had to do with the extent to which the men thought they were getting the pay they deserved and whether they believed it was unfair that certain other people were paid the same.

The next two factors were economic in nature. Factor 4, Total Household Income, consisted of the total (1981) income of all of the members of the men's household, as indicated in response to a direct query, plus the amount of income said to be earned in any part-time work done by household members. Factor 5, on the other hand, did not involve money directly. Based on two job-related questions—whether the fringe benefits and job security were good—Factor 5 was termed Non-Pay Economic Benefits.

Table 2
Factor Correlation Matrix

Factor	1	2	3	4	5	6	7	8
1. Social Comparison Frequency	—	.02	.09	-.02	-.02	.01	-.04	-.05
2. Intrinsic Job Satisfaction		—	-.06	.07	.16	-.20	.16	.23
3. Current Inequity			—	.00	-.05	.05	-.12	-.10
4. Total Household Income				—	.04	-.02	.03	-.05
5. Non-Pay Economic Benefits					—	-.06	.04	.05
6. Satisfaction with Work Environment						—	-.07	-.09
7. Future Equity							—	.10
8. Quality of Life								—

Factor 6 had to do with the respondents' jobs also, but focused on the work environment. The two items loading heavily on this factor were the men's ratings of how pleasant were their coworkers and their job surroundings. Factor 7, again, had to do with equity considerations. Unlike the forementioned Current Inequity factor, however, this particular cluster seemed to be oriented toward the future; the item with the highest loading asked the men to indicate what they thought the chances were that they would be getting the pay they deserve 5 years from now. Finally, Factor 8 was interpreted as a measure of the men's Quality of Life. The three items in this factor asked about (a) the difference between their own and their parents' standard of living, (b) whether they thought their work hours were good, and (c) how satisfied they were with their present standard of living.

In sum, all four proposed psychological determinants of pay satisfaction were reflected in the factors that were isolated from among the 29 predictor items. The economic benefits that the men obtained from their work were tapped by Factors 4 (Total Household Income) and 5 (Non-Pay Economic Benefits), and perhaps Factor 8 (Quality of Life), as well. Equity considerations were involved in Factors 3 (Current Inequity) and 7 (Future Equity). Noneconomic job satisfactions were measured by Factors 2 (Intrinsic Job Satisfaction) and 6 (Satisfaction with Work Environment); and social comparison processes were assessed by Factor 1 (Social Comparison Frequency). Note also that although an oblique rotation was carried out in the factor analysis, the factors have only relatively low correlations with each other, as is shown in Table 2. Intrinsic Job Satisfaction tended to be most strongly related to other factors, but even here, although significant, the correlations were not great: $r_s = .23$ with Quality of Life, $-.20$ with Satisfaction with Work Environment, $.16$ with Non-Pay Economic Benefits, and $.16$ with Future Equity. The other factors generally had much lower intercorrelations.

All 29 items were used in calculating the scores for each respondent on each of the eight factors by weighting the items by their factor score coefficients.

Demographic variables. We sought also to determine whether a number of demographic variables made any significant contributions to the prediction of pay satisfaction. These were (a) age, in years; (b) length of time with the present company or organization, in months; (c) educational level, in terms of eight levels ranging from the noncompletion of high school to the receipt of a doctoral degree; (d) self-assigned social class: upper, middle, working, or lower; (e) occupational level, nine levels from professional to unskilled laborer; (f) member of either union or professional association, or neither; and (g) unemployment experience for at least 1 month: yes or no.

Although we did not have an overall theoretical analysis of the possible influence of these variables, other research (see, e.g., Lawler, 1971), as well as intuition, suggested that they might well contribute to pay satisfaction, and could do so in two different ways. For one, the demographic variables might represent more distal causes that exert their influence on pay satisfaction via more proximate psychological causes.

Significant effects of the demographic variables that are independent of the present eight psychological factors might then indicate that additional psychological causes other than those assessed in this study remain to be identified. Thus, age, length of time with the present organization, and perhaps even prior experience of unemployment might all promote a relatively passive acceptance of one's current employment situation. As a consequence, higher values on these three demographic variables would contribute to greater pay satisfaction independently of the psychological considerations reflected in our factors. Educational level might also operate independently of these factors by decreasing this passive acceptance. The better educated persons could believe that they have more alternatives to their present job and so are less content with their wage, whatever it happens to be. Alternatively, at least some of the demographic variables could lead to pay satisfaction by means of the psychological determinants we have identified. Higher educational and occupational levels and higher social status might promote more frequent and more extensive social comparisons, for example. In engaging in these comparisons fairly readily, then, the people with more education those who are in more prestigious jobs, or those who believe they have higher social status might discover that others earn more than they do, and become dissatisfied. Alternatively, these particular demographic characteristics could conceivably be viewed as "investments" that deserve greater economic returns. High values on these variables, therefore, might lead to a sense of inequity, which, in turn, produces dissatisfaction with one's pay.

These possibilities indicate that we had no single theoretical rationale for specific predictions regarding the relations between the present demographic variables and pay satisfaction. There was one matter about which we were fairly certain, however. In considering the psychological determinants as proximal causes of wage satisfaction and the demographic variables as distal causes, we expected the psychological factors, collectively, to make a far greater contribution to pay satisfaction.

Results

Regression analyses were carried out to determine the extent to which all of the measures previously described contributed to pay satisfaction. We explored first the importance of the psychological factors versus the demographic variables by carrying out a two-step regression analysis in which all eight psychological factors were entered in the initial step and all seven demographic measures were entered next. In this procedure, the former factors yielded an R of .75 ($R^2 = .56$) with pay satisfaction. The addition of the demographic variables in the second step resulted in only a slight improvement in the prediction. The R rose only to .80 ($R^2 = .64$), and the R^2 of the increment was small (.08), although significant, F change (7, 223) = 6.83, $p < .00005$. Reversing the order of entry of these measures led to a

Table 3
Beta Weights of Measures in Prediction of Pay Satisfaction

Measure	β	SE β	T	p
Current Inequity (F3)	-.49	.04	-11.24	<.00005
Intrinsic Job Satisfaction (F2)	.41	.05	8.31	<.00005
Non-Pay Economic Benefits (F5)	.34	.04	8.12	<.00005
Quality of Life (F8)	.30	.05	6.41	<.00005
Future Equity (F7)	.28	.04	6.71	<.00005
Occupational level ^a	.15	.06	2.65	<.01
Age	.14	.06	2.41	<.02
Educational level	-.12	.05	-2.18	<.03
Unemployment experience	-.12	.04	-2.72	<.01
Total Household Income (F4)	.10	.04	2.31	<.02

Nonsignificant predictors

Work Environment (F6)	-.07	.04
Member work association	-.06	.05
Social Comparison Frequency (F1)	-.05	.04
Social class	.003	.05
Time with organization	-.002	.06

Note. $R^2 = .64$; $F(15, 223) = 26.24$, $p < .00005$. F = Factor.

^a Low scores on this variable indicate higher occupational status.

similar picture. In this case, the seven demographic variables in the first step gave rise to a significant R of .48 ($R^2 = .23$), but the inclusion of the eight factors in the next step improved the prediction of pay satisfaction greatly: the R , obviously, was again .80, but the R^2 of the increment was now .41, F change (8, 223) = 31.41, $p < .00005$. All in all, it appears that the psychological factors were more important than our demographic variables, and it may be that the latter tended to affect pay satisfaction largely, although not solely, through the psychological considerations assessed by the factors.

More detailed information about the relative contributions made by each of our measures can be obtained from Table 3, which summarizes the regression analysis involving all 15 variables. Table 3 lists the resulting beta weights in the prediction of pay satisfaction, with their standard errors and significance levels.³

Table 3 clearly shows that five of the eight psychological factors contributed substantially to the prediction of pay satisfaction and, furthermore, that these factors tapped three of the four classes of psychological determinants we have proposed. The men's beliefs as to whether their wages were fair seemed to be the most important of all. They were especially apt to have relatively little satisfaction with their pay if they were high on Current Inequity ($\beta = -.49$) but, on the other hand, tended to be happier with their present earnings if they thought there was a good chance that they would get the pay they deserved in the future ($\beta = .28$). Intrinsic Job Satisfaction was somewhat less important than Current Inequity ($\beta = .41$) but contributed rather more to the prediction of our criterion measure than did either of the economically oriented indices separately: Non-Pay Economic Benefits ($\beta = .34$) and Quality of Life ($\beta = .30$). Interestingly, the factor based most clearly and most directly on money, Total Household Income, was the least important of all our predictors ($\beta = .10$). It is also worth pointing out that the Social Comparison Frequency factor did not make a significant contribution to the prediction equation at all.

In accord with the previously reported exploratory analyses, the demographic variables were generally less important than our psychological factors; only four of the seven demographic items contributed significantly to the prediction of pay satisfaction, and their beta weights were fairly small. Age, as expected, was positively related ($\beta = .15$), but occupational level had a negative relation with the criterion ($\beta = .14$); men in the higher status occupations tended to be less satisfied with their pay. Educational level and unemployment experience were also negatively related to pay satisfaction (both β s = $-.12$).

Discussion

Psychological Determinants

Although the present findings are somewhat unexpected in several respects, in other ways they are consistent with published research and theory. Most notably, the results lend some support to prevailing thinking about the importance of perceived equity as a determinant of pay satisfaction (e.g., Lawler, 1971). The more strongly the men believed they were receiving the pay they deserved, and the smaller the discrepancy between their reported income and the income they thought they deserved relative to their actual pay, the more satisfied they were with their earnings.

The regression analysis also indicates that a sense of equity was a stronger predictor of pay satisfaction than were the material benefits. However, the factor analysis raises a question about such a simple assertion. Yearly pay was sufficiently strongly correlated with the individual equity items to load strongly ($-.59$) on the Current Inequity factor. This suggests that equity considerations are not wholly independent of actual pay, even if in principle they might be. It is, in fact, misleading to pit material benefits against nonmaterial feelings of equity, as if they were mutually exclusive. Our study shows that, in part, they are related and that, in part, they both make substantial and independent contributions to pay satisfaction.

One respect in which Lawler's (1971) model needs qualifying is that it assumes that all potential determinants of pay satisfaction operate via equity calculations. The most immediate determinant of pay satisfaction supposedly is a comparison between one's deserved and one's actual pay. By contrast, our findings reveal that other considerations, such as the material benefits obtained through one's earnings and job satisfaction, make independent contributions toward pay satisfaction.

Intrinsic job satisfaction is a third important but apparently neglected predictor of pay satisfaction. The extent to which that relation depends on a conscious trade-off of the two sets of benefits or on the direct generalization of affect—as discussed in the beginning of this article—remains to be studied.

One surprising finding has been the consistent failure of social comparisons to play any significant part in directly predicting pay satisfaction. As we stated, although most social psycho-

³ The possibility cannot be excluded that similarities and differences in item formats may have made some contribution to these findings. Note, however, that all psychological items did not have formats similar to the items assessing pay satisfaction, and among psychological items with similar formats several different factors were identified.

logical discussions in this area maintain that income satisfaction is directly affected by social comparisons, some analyses suggest that satisfaction with this outcome need not be influenced by such comparisons. If accepted at face value, our findings indicate that the taken-for-granted centrality of comparisons with other individuals or groups merits a closer, more critical consideration. A person could determine how well off he or she is in the world in ways other than through the conventionally assumed social comparisons, particularly outside the laboratory. In the case of pay, one might evaluate his or her earnings by judging them against expectations based on earlier income (Messé & Watts, 1983) or, as Goodman (1974) has pointed out, on company policies or self-conceptions. For that matter, as we have proposed, satisfaction with income might even be influenced by feelings toward other aspects of work through a generalization process. After a delicious meal, the world can seem a very satisfying place, particularly if the good food is accompanied by good wine; it is quite unnecessary to check that one's own meal was better than anyone else's. But even when social comparisons do affect wage satisfaction, they may do so through comparisons that are not usually considered in the pay-satisfaction literature. According to our results, a man might match the standard of living his earnings now bring him with the standard of living his parents had experienced when he was a child.

On the other hand, conventional social comparisons should not be too readily ignored. Our assessment of social comparisons consisted of a series of questions about how frequently comparisons were engaged in. We did not establish what individuals discovered about their pay compared with that of others or how they felt about what they learned. It may well be that a more elaborate analysis of comparison processes would reveal that social comparisons are related to pay satisfaction because they determine either feelings of equity or, alternatively, judgments of material well-being and standard of living.

What is clear from our study is that almost two thirds of the variance in direct expressions of pay satisfaction from a random sample of full-time male workers can be predicted from measures of equity, material benefits, and intrinsic job satisfaction, together with a handful of simple demographic variables.

Demographic Variables

Our findings suggest that four demographic variables—age, education, occupational level, and experience with unemployment—improve the prediction of pay satisfaction beyond that obtained with the psychological variables, although only to a relatively small extent. One of these relations is fairly straightforward. As might be expected, older respondents were more satisfied with their pay. The other results are somewhat more surprising. Higher educational level was associated with lower satisfaction with pay, a finding that has been noted elsewhere (Klein & Maher, 1966). And moreover, although occupational level is significantly correlated with educational level ($r = .59$), it made its own significant contribution to the prediction of pay satisfaction, and in the same direction as educational level. Men at the higher levels were more dissatisfied with their pay than were the men at the lower levels. Because neither educational nor occupational levels were correlated with the Current Inequity factor, these variables evidently did not contribute to pay

satisfaction through equity considerations. It seems more likely, we believe, that they operated through a process we had not anticipated at the start of our research. The respondents might have thought that higher educational and occupational levels increased their job opportunities so that those at the higher levels did not have to accept their current pay in a relatively passive manner. They could look elsewhere for employment.

Another somewhat unexpected predictor was past experience of unemployment. Psychologically, one might have expected unemployment experience to be positively associated with pay satisfaction, on the assumption that any pay is more satisfying than no pay. In fact, our finding was the opposite: Those with past experience of unemployment were less satisfied with their current pay.

Our assumption is that, in general, demographic variables are distal influences on individuals' actions and experiences and they exert their influence through proximate influences captured by psychological variables. The fact that four demographic variables improved, albeit modestly, the prediction of pay satisfaction suggests to us that some psychological variables, in addition to those we have studied, remain to be identified before a complete account of pay satisfaction can be offered.

References

- Adams, J. S. (1965). Injustice in social exchange. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 267–299). New York: Academic Press.
- Andrews, I. R., & Henry, M. M. (1963). Management attitudes towards pay. *Industrial Relations*, 3, 29–39.
- Clore, G. L., & Byrne, D. (1974). A reinforcement affect model of attraction. In T. L. Huston (Ed.), *Foundations of interpersonal attraction* (pp. 143–170). New York: Academic Press.
- Crosby, F. (1976). A model of egoistical relative deprivation. *Psychological Review*, 83, 85–113.
- Davies, J. C. (1962). Toward a theory of revolution. *American Sociological Review*, 27, 5–19.
- Dyer, L., & Theriault, R. (1976). The determinants of pay satisfaction. *Journal of Applied Psychology*, 61, 596–604.
- Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, 7, 117–140.
- Finn, R. H., & Lee, S. M. (1972). Salary equity: Its determination, analysis, and correlates. *Journal of Applied Psychology*, 56, 283–292.
- Goodman, P. S. (1974). An examination of referents used in the evaluation of pay. *Organizational Behavior and Human Performance*, 12, 170–195.
- Goodman, P. S., & Friedman, A. (1971). An examination of Adams' theory of inequity. *Administrative Science Quarterly*, 16, 271–288.
- Gurr, T. R. (1970). *Why men rebel*. Princeton, NJ: Princeton University Press.
- Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.
- Hyman, H. H. (1942). The psychology of status. *Archives of Psychology*, Columbia University, No. 269.
- Isen, A. M. (1984). Toward understanding the role of affect in cognition. In R. Wyer & T. Srull (Eds.), *Handbook of social cognition* (pp. 179–236). Hillsdale, NJ: Erlbaum.
- Jaques, E. (1961). *Equitable payment*. New York: Wiley.
- Klein, S. M., & Maher, J. R. (1966). Education level and satisfaction with pay. *Personnel Psychology*, 19, 195–208.
- Lawler, E. E. (1971). *Pay and organizational effectiveness: A psychological view*. New York: McGraw-Hill.

- Leventhal, G. S., & Anderson, D. (1970). Self-interest and the maintenance of equity. *Journal of Personality and Social Psychology*, 15, 57-62.
- Martin, J., & Murray, A. (1983). Distributive injustice and unfair exchange. In D. M. Messick & K. S. Cook (Eds.), *Equity theory: Psychological and sociological perspectives*. New York: Praeger.
- Messé, L. A., & Watts, B. L. (1983). Complex nature of the sense of fairness: Internal standards and social comparison as bases for reward evaluations. *Journal of Personality and Social Psychology*, 45, 84-93.
- Motowidlo, S. J. (1982). Relationship between self-rated performance and pay satisfaction among sales representatives. *Journal of Applied Psychology*, 67, 209-213.
- Patchen, M. (1961). *The choice of wage comparisons*. Englewood Cliffs, NJ: Prentice-Hall.
- Pettigrew, T. (1967). Social evaluation theory: Convergences and applications. In D. Levine (Ed.), *Nebraska Symposium on Motivation* (pp. 241-311). Lincoln: University of Nebraska Press.
- Ronan, W. W., & Organt, G. J. (1973). Determinants of pay and pay satisfaction. *Personnel Psychology*, 26, 503-520.
- Ronen, S. (1986). Equity perception in multiple comparisons: A field study. *Human Relations*, 39, 333-346.
- Runciman, W. G. (1966). *Relative deprivation and social justice*. London: Routledge & Kegan Paul.
- Schwab, D. P., & Wallace, M. J., Jr. (1974). Correlates of employee satisfaction with pay. *Industrial Relations*, 13, 78-89.
- Thibaut, J. W., & Kelley, H. H. (1959). *The social psychology of groups*. New York: Wiley.
- Weiner, N. (1980). Determinants and behavioral consequences of pay satisfaction: A comparison of two models. *Personnel Psychology*, 33, 741-757.
- Weitzel, W., Harpaz, I., & Weiner, N. (1977). Predicting pay satisfaction from nonpay work variables. *Industrial Relations*, 16, 322-334.

Received October 21, 1985

Revision received March 25, 1987

Accepted May 4, 1987 ■

Schmitt Appointed Editor, 1989-1994

The Publications and Communications Board of the American Psychological Association announces the appointment of Neal Schmitt, Michigan State University, as editor of the *Journal of Applied Psychology* for a 6-year term beginning in 1989. As of January 1, 1988, manuscripts should be directed to

Neal Schmitt
Department of Psychology
Psychology Research Building
Michigan State University
East Lansing, Michigan 48824

Manuscript submission patterns for the *Journal of Applied Psychology* make the precise date of completion of the 1988 volume uncertain. The current editor, Robert Guion, will receive and consider manuscripts until December 31, 1987. Should the 1988 volume be completed before that date, manuscripts will be redirected to Schmitt for consideration in the 1989 volume.

Employee Age as a Moderator of the Relation Between Perceived Work Alternatives and Job Satisfaction

Samuel B. Pond III and Paul D. Geyer
North Carolina State University

This study was designed to explore whether employee age influences the relation between perceived work alternatives and job satisfaction. Moderated regression analyses were conducted using the survey responses of 226 employees between the ages of 24 and 50 who worked for a mental health institution. The analyses revealed that a Perceived Work Alternatives \times Employee Age interaction significantly predicted job satisfaction. Neither organizational tenure nor employee educational level accounted for job-satisfaction variance beyond that accounted for by perceived work alternatives alone, nor did they interact with perceived work alternatives to predict job satisfaction. These findings indicate that employee age is associated with the relation between perceived alternatives and job satisfaction. They also provide some insight into which of a number of age-related effects may be most pertinent to this relation.

Although the presence of an inverse relation between perceived work alternatives and job satisfaction is well established in the literature, an unequivocal understanding of this relation does not exist (Hulin, Roznowski, & Hachiya, 1985; Miller, Katerburg, & Hulin, 1979; Mobley, Horner, & Hollingsworth, 1978). It is possible that perceived work alternatives and job satisfaction covary because of some other variables (e.g., personal characteristics of the employee). It is also possible that the relation is a causal one, so that an increase or decrease in job satisfaction leads to a respective decrease or increase in beliefs about the availability of better work alternatives. Most researchers, however, strongly suggest that the perception of the availability of better (and poorer) work alternatives directly affects job satisfaction (e.g., Hulin et al., 1985; Miller et al., 1979; Smith, Kendall, & Hulin, 1969).

Even if one accepts that the perception of better work alternatives directly affects job satisfaction, however, one probably should not accept the notion that the perception has the same impact on all employees (Hulin et al., 1985). Therefore, it becomes necessary to use other variables to clarify the nature of the relation. This study focuses on employee age as a possible moderator of the relation between perceived alternatives and job satisfaction.

In order to aid in the clarification of the relation between perceived work alternatives and job satisfaction, an adaptation cycle model presented by Rosse & Miller (1984, pp. 207-215) is introduced in Figure 1. Rosse and Miller's model describes how employees consider a number of alternative strategies to reduce (or "fix") the "relative dissatisfaction" occasioned by a "stimulus event." Four factors (i.e., personal experience, role model-

ing, social norms, and perception of constraints) are hypothesized to influence which alternative strategies are developed and how they are assessed for use. The strategy with the highest perceived utility is used. If it does not work, another cycle begins and new strategies are used.

Although this model has been used to show how, why, and when withdrawal behaviors (absenteeism, turnover, lateness, etc.) might develop among employees, we believe that it can also provide a useful theoretical framework for thinking about the relation between perceived work alternatives and job satisfaction. For example, by using Rosse & Miller's model, the perception of better work alternatives can be portrayed as a stimulus event that might bring on relative dissatisfaction.

Critical to the thesis of this study is that the model provides a theoretical basis for why an employee characteristic, such as age, might moderate the strength of the relation between perceived work alternatives and job satisfaction. Four possibilities associated with employee age are suggested here, all of which indicate that a weaker correlation between perceived work alternatives and job satisfaction should exist for older employees relative to younger employees. For simplicity of presentation, all of these possibilities are presented from the perspective of the older employee.

First, older employees, relative to younger employees, may not be bothered *as much* by their perceptions of better work alternatives. Change in values, higher status, higher pay, more prestige, and investments in their current job may not cause perceived work alternatives to be a major stimulus event for older employees (Hall & Mansfield, 1975; Rhodes, 1983).

Second, it may be that older employees are bothered by their perceptions of better work alternatives but are more able to deal with this aggravating stimulus event by virtue of their job position, tenure, and so forth. They may have more power to effect changes in their job situation so that perceived deficiencies in their jobs can be corrected (Hulin et al., 1985).

Third, it may be that the aspirations of older employees become "ground down" over time so that these employees become more resigned to their job situation and less likely to consider

We wish to express our sincere appreciation to the two anonymous reviewers who worked very hard and provided us excellent guidance throughout the revision process.

Paul Geyer is currently at Appalachian State University.

Correspondence concerning this article should be addressed to Samuel B. Pond, III, Department of Psychology, North Carolina State University, Box 7801, Raleigh, North Carolina 27695-7801.

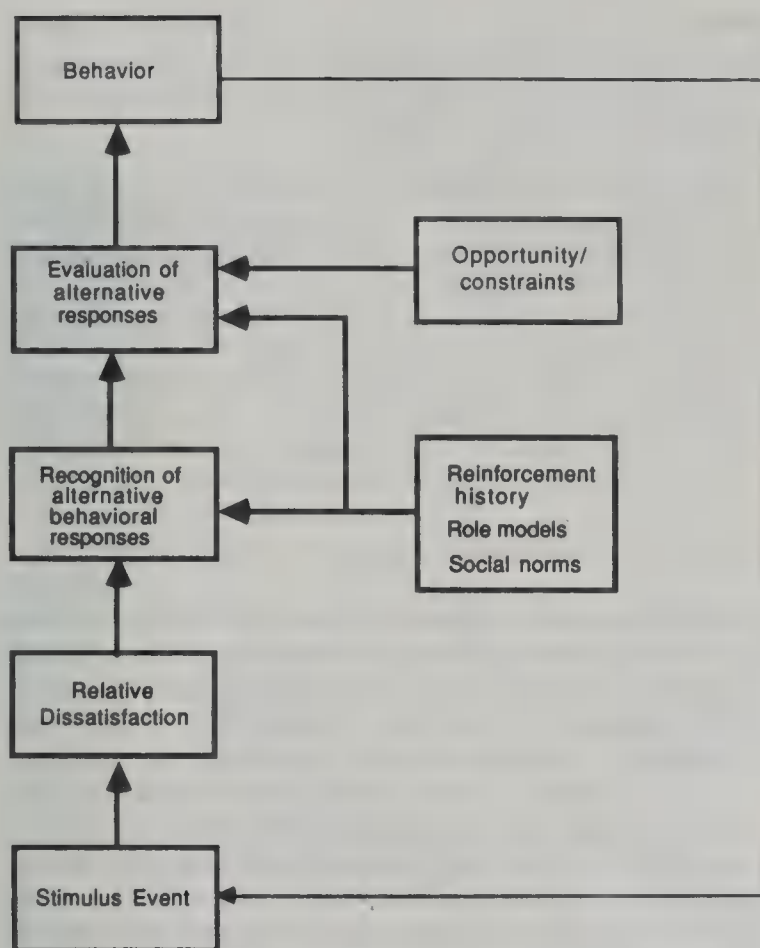


Figure 1. Model of the adaptation cycle presented by Rosse & Miller. (From *Absenteeism*, p. 208, edited by P. S. Goodman and R. S. Atkins, 1984, San Francisco, CA: Jossey-Bass. Adapted by permission.)

perceived work alternatives as a major stimulus event (Seashore, 1974; Wright & Hamilton, 1978). Also, these employees may be aware of the difficulties that older employees often face when they seek new employment.

Finally, older employees, relative to younger employees, may not be bothered *as much* by their perceptions of better work alternatives because of certain cohort effects. For example, older employees as a group, given their common history, may differ from the younger employees on which job factors they consider to be important determinants of certain job attitudes or beliefs (e.g., attitudes about commitment to a company; Riley & Foner, 1968). For these and other similar reasons, perceived work alternatives may not enter as strongly into the frame of reference of older employees when they are judging job satisfaction (Smith et al., 1969).

It is apparent from these explanations that age represents a great deal about an employee. As Rhodes (1983) noted, age is a variable that is associated with an interrelated group of effects that influences work attitudes and behavior: (a) psychosocial and biological aging effects, (b) cohort effects, and (c) period effects (interpreted in this study as work-experience effects).

Psychosocial and biological aging are described by Rhodes (1983) as developmental components of an overall "aging effect" (what she has also termed *chronological age*). Psychosocial aging refers to "systematic changes in personality, needs, expectations, and behavior as well as performance in a sequence of socially prescribed roles" (p. 329). Rhodes provided evidence

that employees' work values, needs, and outlooks change as employees grow older. Promotion opportunities and retirement benefits, for example, are often valued differently by people depending on which career stage they are in (Schein, 1978; Super, 1980). Biological aging includes changes in a person's sensorimotor performance, balance, and so forth. As employees grow older, the importance of the physical demands of various job alternatives are often weighted differently.

Cohort effects, another influence on work attitudes and behaviors associated with employee age, refer to the unique behavior of different age groups because of such things as different amounts of education or common life experiences. As an example of a possible cohort effect on the relation between perceived work alternatives and job satisfaction, consider the following: Younger employees, relative to their older co-workers have, as a group, grown up in a time in which there has been little, if any, stigma attached to changing jobs either for the sake of advancement or just for the sake of change (Arnold & Feldman, 1986; Whyte, 1957). Consequently, younger and older employee cohorts might value job mobility differently (Schein, 1978). This difference, then, could alter the relation between perceived work alternatives and job satisfaction among employees of different age groups.

Finally, it is reasonable to assume that employee age may influence the relation between perceived work alternatives and job satisfaction because of the work-experience effects associated with it. The work-experience effects embodied by employee age pertain to the impact that learning about a job and the job environment has on an individual. As noted by Rhodes (1983), employees' views regarding aspects of work such as co-worker relations, supervision, reward structure, and market conditions are subject to temporal changes. Not only does work experience affect employee perceptions of job alternatives and job satisfaction, but it also determines how people will respond to their perceived work alternatives (Hulin et al., 1985; Mobley, 1982; Wanous, 1980).

Because employee age embodies so much information about a person, Rhodes (1983) has endorsed multivariate research that attempts to examine the complex relation between age and job-related variables. In the study described here, an employee-education variable is used to help probe that part of employee age associated with cohort effects. A large literature exists in support of using educational level as a proxy for these effects (Rhodes, 1983; Riley & Foner, 1968). Similarly, an organizational-tenure variable is included to help examine the effects of work experience (Rhodes, 1983).

If employee age moderates the relation between perceived work alternatives and job satisfaction, exploring the interrelations among employee age, educational level, and organizational tenure may provide some more insight into the nature of this moderating effect. This could, subsequently, lead to a more refined exploration of the impact of employee age on the relation between perceived work alternatives and job satisfaction and of its role in other organizational research issues.

Method

Subjects

The sample consisted of 226 employees working in a mental health institution located in the southeastern United States. Approximately

24% of the employees were men, and about 76% were White. Sixty-seven percent of the employees participating in the study were married, and most of them had college degrees. The median organizational tenure for employees in this sample was in the range of 6 to 10 years.

For practical considerations, data were collected from employees of all ages. However, only data from employees 25 to 49 years of age were used in this study. This sample incorporated about 80% of the work force and avoided the extremely young and old employees. Employee ages were grouped into units of 5 years and the sample was distributed across these groupings as follows: 25 to 29 years ($n = 59$), 30 to 34 years ($n = 65$), 35 to 39 years ($n = 55$), 40 to 44 years ($n = 25$), and 44 to 49 years ($n = 22$). This age distribution is very similar to that of the U.S. work force in general (U.S. Department of Labor, 1986).

The literature indicates that the relation between job satisfaction and age can be quite variable in terms of direction and strength for employees younger than 25 years of age and for employees considering retirement (Gibson & Klein, 1970; Herzberg, Mausner, Peterson, & Capwell, 1957; Saleh & Otis, 1964) for reasons that are quite likely to be specific to these employees (Super, 1980). In the organization studied here, it was reasonable to assume that employees 50 years and older would be thinking seriously about retirement.

Measures

Job satisfaction. Job satisfaction was measured using a modified version of Quinn and Shepard's (1974) general job-satisfaction scale. Each of the respondents was asked to select one of five response categories which best answered each of six questions: (a) "If you had to decide all over again whether to take the job you now have, what would you decide?" (1 = *definitely not take job*, 5 = *definitely take job*); (b) "If a friend asked if he/she should apply for a job like yours with your employer, what would you recommend?" (1 = *recommend not at all*, 5 = *recommend strongly*); (c) "How does this job compare to your ideal job?" (1 = *very far from ideal*, 5 = *very close to ideal*); (d) "How does your job measure up to the sort of job you wanted when you took it?" (1 = *not at all like what I wanted*, 5 = *just like what I wanted*); (e) "All things considered, how satisfied are you with your current job?" (1 = *not at all satisfied*, 5 = *completely satisfied*); and (f) "In general, how much do you like your job?" (1 = *not at all*, 5 = *a great deal*).

A mean of the six items served as a job satisfaction score for each employee. Low and high scores represented low and high levels of job satisfaction, respectively. A coefficient alpha reliability estimate of .90 was obtained for this scale.

Perceived work alternatives. Employee beliefs about the availability of better work alternatives were assessed by asking the following: "How likely is it that you could get a job as good as yours but with (a) better pay, (b) nicer co-workers, (c) more satisfactory supervision, (d) more chances for advancement, (e) more interesting work, (f) better working conditions, (g) more job security, (h) more meaningful work?" The eight work-alternative items used in this study were developed by Farrell and Rusbult (personal communication, September 1983). We modified the response format, however, to include four options instead of three.

A four-point response scale was used with options ranging from *very unlikely* (1) to *very likely* (4). For each employee, a mean of the eight items served as a perceived work-alternatives score. Employees scoring high on this scale believed that better work alternatives were available; low scores indicated the opposite. A coefficient alpha reliability estimate of .84 was obtained.

Demographic information. Six questions were asked so that the following information could be obtained for each employee: age (1 = 25 to 29 years, 2 = 30 to 34 years, 3 = 35 to 39 years, 4 = 40 to 44 years, 5 = 45 to 49 years); sex (0 = male, 1 = female); race (0 = Black, or "other", 1 = White); marital status (0 = not married, 1 = married); educational level (1 = high school diploma or below, 2 = 1 to 3 years of college, 3 = college degree, 4 = post-baccalaureate degree); and organiza-

tional tenure (1 = less than 1 year, 2 = 1 to 2 years, 3 = 3 to 5 years, 4 = 6 to 10 years, 5 = 11 years or more).

Questionnaire Administration

Questionnaires were administered by university personnel to groups of employees (*Mdn* group size = 15 respondents) during routinely scheduled meetings and at meetings held specifically for questionnaire administration. At each of the meetings, envelopes containing cover letters, consent forms, instructions, and questionnaires were distributed to the employees. Employees were informed of the nature of the research and the reasons why they were selected and were reminded both that participation was voluntary and that any information they provided would be confidential. Upon completing the questionnaires, employees returned them directly to university personnel.

Results

Variable means, standard deviations, and intercorrelations (with their significance levels) are presented in Table 1. Consistent with the findings of Miller et al. (1979) and Mobley et al. (1978), the perceived work alternatives variable correlated significantly and negatively with job satisfaction ($r = -.49$) and accounted for approximately 24% of the job-satisfaction variance. Also, consistent with Rhodes's (1983) review of the literature, it was found that age, organizational tenure, and educational level correlated significantly with one another. Employee age correlated positively with organizational tenure ($r = .31$) and negatively with education ($r = -.21$). None of these variables correlated significantly with job satisfaction, and only age correlated significantly with perceived work alternatives ($r = .11$).

The prediction of job satisfaction with perceived work alternatives was enhanced, however, when employee age was combined with perceived work alternatives in a regression formula. As shown in Table 2, combining perceived work alternatives, employee age, and their interaction yielded a multiple correlation of .53 (adjusted R of .51). Together these variables accounted for about 28% of the total job satisfaction variance, $F(3, 222) = 28.20, p < .01$.

As shown in Table 2, employee age significantly accounted for about an additional 2% of variance over that attributed to perceived work alternatives, and, most notably, the Perceived Work Alternatives \times Age interaction significantly accounted for about 2% more variance over this. The perception of work alternatives was related differently with job satisfaction depending on an employee's age.

A graph of the Perceived Work Alternatives \times Employee Age interaction appears in Figure 2. A single regression formula was used to derive each of the regression lines shown. (See Cohen & Cohen, 1975, for a detailed explanation of this approach.) For purposes of clarity, only lines for the two extreme age groups have been presented. The slopes of the lines for the other three age groups would fall within the slopes of these two lines.

For illustrative purposes, then, a value of 1 was used in the regression formula to represent the age variable in the line shown for the youngest group of employees (25- to 30-year-olds). The value of 5 was used to derive the line representing the oldest employee age group (45- to 49-year-olds). When the slopes of the two lines are compared, it is apparent that there is a stronger negative relation between perceived work alternatives

Table 1
Means, Standard Deviations, and Intercorrelations of Job-Satisfaction, Perceived Work-Alternatives, Age, Tenure, and Education Variables

Variable	M	SD	1	2	3	4	5
1. Job satisfaction	3.49	0.86	(.90)				
2. Perceived work alternatives	2.64	0.58	-.49**	(.84)			
3. Employee age	2.50	1.26	.07	.11*	—		
4. Tenure	3.50	1.27	-.01	-.09	.31**	—	
5. Education	2.55	0.96	-.08	.03	-.21**	-.37**	—

Note. $N = 226$. Coefficient alpha reliability estimates are reported in parentheses. See Demographic Information section for coding of age, tenure, and education.

* $p < .05$. ** $p < .01$.

and job satisfaction for the younger employees than for the older employees.

In spite of its low zero-order correlation with job satisfaction, this regression analysis showed that employee age interacted with perceived work alternatives to explain more job-satisfaction variance. This did not occur when the organizational-tenure or employee-education variables were used in place of employee age in similar regression analyses (see Table 2).

Because employee age interacted with perceived work alternatives to predict job satisfaction, one final moderated regression analysis was conducted to try to identify which of the age-related effects discussed earlier might be most directly associated with the moderating effect of age on the relation between perceived work alternatives and job satisfaction. In order to do this, job-satisfaction variance associated with perceived work alternatives and with organizational tenure and education (the latter two were entered into the regression equation as a set) was removed to see if employee age and the Perceived Work Alternatives \times Employee Age interaction would still account for a significant amount of the remaining job-satisfaction variance. The results of this analysis are shown in Table 3.

A multiple correlation of .54 (adjusted R of .52) was found, and both employee age and the Perceived Work Alternatives \times Employee Age interaction made a significant and unique contribution to the prediction of job satisfaction, $F(1, 220) =$

17.80, $p < .01$. The combination of age and its interaction with work alternatives accounted for about an additional 4% of variance.

Discussion

The negative correlation between perceived work alternatives and job satisfaction found in other studies was also found in this study. However, the moderating effect of employee age suggests that the relation between perceived work alternatives and job satisfaction is weaker for older than for younger employees. Furthermore, the results of this study imply that some aspect of employee age other than that related to organizational tenure

Table 2
Results of Moderated Regression Analyses

Variable	R^2	ΔR^2	df	F (step)
Perceived alternatives (PA)	.241	.241	1, 224	70.94**
Employee age (EA)	.257	.016	1, 223	4.84*
EA \times PA	.276	.019	1, 222	5.92*
Perceived alternatives	.241	.241	1, 224	70.94**
Tenure (T)	.243	.002	1, 223	0.77
T \times PA	.243	.000	1, 222	0.98
Perceived alternatives	.241	.241	1, 224	70.94**
Education (E)	.245	.004	1, 223	1.21
E \times PA	.245	.000	1, 222	0.17

Note. $N = 226$.

* $p < .05$. ** $p < .01$.

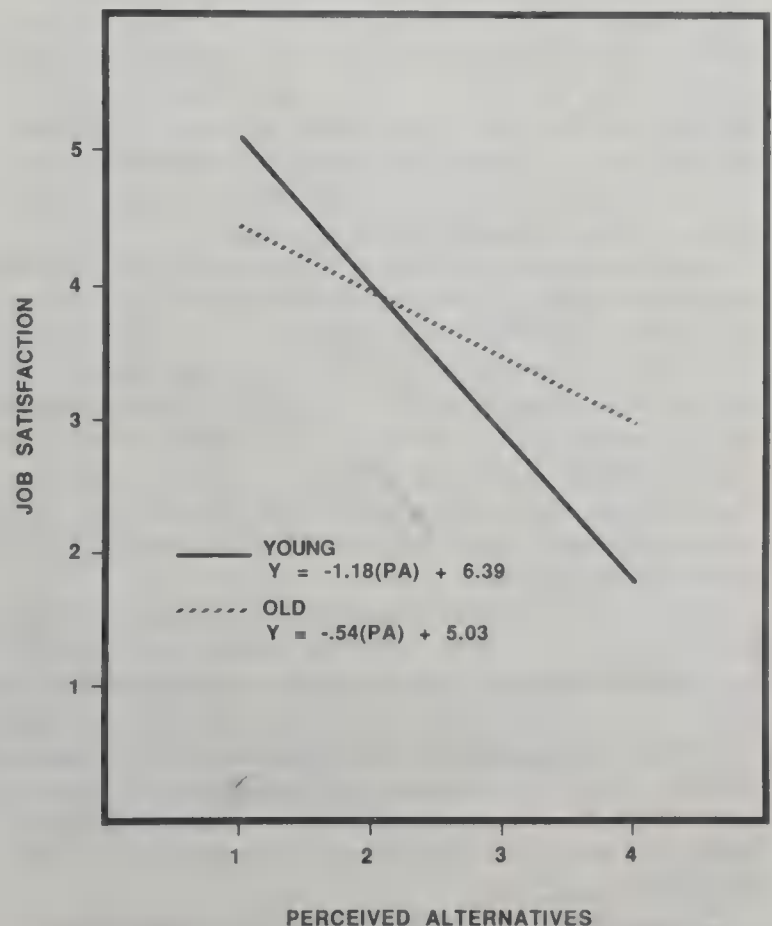


Figure 2. Interaction effect of perceived work alternatives and employee age on job satisfaction.

Table 3
Moderated Regression Results Using Tenure, Education, Age, and Perceived Work Alternatives to Predict Job Satisfaction

Step	Variable	R ²	ΔR ²	df	F (step)
1	Perceived alternatives (PA)	.241	.241	1, 224	70.94**
2	Education and tenure	.251	.011	1, 222	1.56
3	Employee age (EA)	.271	.020	1, 221	6.15*
4	EA × PA	.288	.017	1, 220	5.14*

Note. N = 226.
* p < .05. ** p < .01.

and employee educational level is probably responsible for its moderating effect.

It was presumed that variance associated with cohort and work-experience effects would be (to at least some extent) represented by the employee educational-level and organizational-tenure variables, respectively. Although these variables were both significantly correlated with employee age, neither of them contributed to the prediction of job satisfaction even when they were combined with the perceived work-alternatives variable. Furthermore, employee age and its interaction with perceived work alternatives continued to account for a significant amount of job-satisfaction variance even after variance associated with organizational tenure and employee educational level had been removed first.

Even though the organizational-tenure variable measured in this study did not correlate highly with job satisfaction or perceived work alternatives, the importance of work-experience effects on the relation between perceived work alternatives and job satisfaction should not be discounted. Organizational tenure measured alone does not accurately reflect work experience received in other organizations. It is also possible that the range of organizational tenure was restricted because of range restrictions intentionally imposed on employee age.

Thus, although not conclusively ruling out cohort effects, the results of this study suggest that the effects of both psychosocial and biological aging as well as that of work experience on the relation between perceived work alternatives and job satisfaction should be closely examined. It is possible that the recognition of perceived work alternatives as a significant stimulus event varies among individuals because of these aspects of age. The development of strategies used to cope with whatever “relative dissatisfaction” might result could also vary among individuals for similar reasons.

Rhodes (1983) and others (e.g., Hall & Mansfield, 1975) have pointed out that the values, needs, and expectations of employees change as they go through different developmental stages in their lives and careers. If employees in different career stages differ with respect to values, needs, and expectations, it is likely that they value outside job opportunities differently. This could help explain the difference in correlations that was found between perceived work alternatives and job satisfaction within the different age groups.

Changes in expectations can be explained not only by psychosocial and biological aging, but also by the work experience employees receive. Differences in the strength of the relation be-

tween perceived work alternatives and job satisfaction among younger and older workers may be the result of these employees acquiring differing amounts and kinds of work experiences. Thus, for example, their response to perceived work alternatives may vary according to the amount of subjective and objective investment employees have in a job (Farrell & Rusbult, 1981; Hall, 1976; Rusbult & Farrell, 1983), or the amount of control employees believe they have over the job situation (Seashore, 1974). Hulin et al. (1985) and Seashore (1974) noted that older workers, by virtue of their age, tenure, and likely higher position in the organization, may have the power to control their work situation by changing the company rules. Younger workers, who usually have less control over their jobs, may be more apt to leave them, especially if they believe better alternatives are available.

Although these interpretations are offered to help understand why employee age should interact with perceived work alternatives to predict job satisfaction, we point out that this exploratory field study involved employees from one work setting (a mental health facility) and that generalization of these results to employees in different work settings should be carefully considered. We also caution that the use of the terms *older* and *younger* in this article must be considered relative to the age range of the employees included in our sample.

It should also be noted that employee age and its interaction with perceived work alternatives did not account for a very large amount of job-satisfaction variance. This is probably because employee age, by itself, is a nonspecific proxy variable for many other important age-related variables, such as those that have been suggested in this study. Also, the low correlation between job satisfaction and employee age is not so surprising because, as Rhodes (1983) has pointed out, the correlations found between age and job attitudes in general are usually not very high—even though age is the most consistent and strongest predictor of job attitudes and behavior of all the demographic variables.

These limitations aside, this study gives insight into the relation between perceived work alternatives and job satisfaction. Specifically, this study implies that in that relation, work experience and psychosocial and biological-aging effects are further relevant aspects of employee age to consider. This finding may prove useful to researchers examining those situations in which employees must choose a strategy that allows them to deal with perceptions of highly desirable work alternatives (Hulin et al., 1985).

More research will be needed to address how employees of different ages and in different stages of their careers actually conceptualize perceived work alternatives and job satisfaction. Future investigation of the moderating effect of employee age on the relation between perceived work alternatives and job satisfaction may benefit from testing implications drawn from models of life and career stages (Levinson, Darrow, Klein, Levinson, & McKee, 1974; Rush, Peacock, & Milkovich, 1980; Super, 1980).

References

Arnold, H. J., & Feldman, D. C. (1986). *Organizational behavior*. New York: McGraw-Hill.
Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. New York: Wiley.

- Farrell, D., & Rusbult, C. E. (1981). Exchange variables as predictors of job satisfaction, job commitment, and turnover: The impact of rewards, costs, alternatives, and investments. *Organizational Behavior and Human Performance*, 27, 78-95.
- Gibson, J. L., & Klein, S. M. (1970). Employee attitudes as a function of age and length of service: A reconceptualization. *Academy of Management Journal*, 13, 411-425.
- Hall, D. T. (1976). *Careers in organizations*. Pacific Palisades, CA: Goodyear.
- Hall, D. T., & Mansfield, R. (1975). Relationships of age and seniority with career variables of engineers and scientists. *Journal of Applied Psychology*, 60, 201-210.
- Herzberg, F., Mausner, B., Peterson, R. O., & Capwell, D. F. (1957). *Job attitude research and opinion*. Pittsburgh, PA: Psychological Service of Pittsburgh.
- Hulin, C. L., Roznowski, M., & Hachiya, D. (1985). Alternative opportunities and withdrawal decisions: Empirical and theoretical discrepancies and an integration. *Psychological Bulletin*, 97, 233-250.
- Levinson, D. J., Darrow, C., Klein, E., Levinson, M., & McKee, B. (1974). The psychological development of men in early adulthood and the mid-life transition. In D. F. Ricks, A. Thomas, & M. Roff (Eds.), *Life history research in psychopathology* (Vol. 3, pp. 243-258). Minneapolis: University of Minnesota Press.
- Miller, H. E., Katerberg, R., & Hulin, C. L. (1979). Evaluation of the Mobley, Horner, and Hollingsworth model of employee turnover. *Journal of Applied Psychology*, 64, 509-517.
- Mobley, W. H. (1982). *Employee turnover: Causes, consequences, and control*. Reading, MA: Addison-Wesley.
- Mobley, W. H., Horner, S. D., & Hollingsworth, A. T. (1978). An evaluation of precursors of hospital employee turnover. *Journal of Applied Psychology*, 63, 408-414.
- Quinn, R. P., & Shepard, L. (1974). *The 1973-1974 quality of employment survey: Descriptive statistics*. Ann Arbor, MI: Institute for Social Research, Survey Research Center.
- Rhodes, S. R. (1983). Age-related differences in work attitudes and behavior: A review and conceptual analysis. *Psychological Bulletin*, 93, 328-367.
- Riley, M. W., & Foner, A. (1968). *Aging and society*. New York: Russell Sage Foundation.
- Rosse, J. G., & Miller, H. E. (1984). An adaptation cycle interpretation of absence and withdrawal. In P. S. Goodman & R. S. Atkins (Eds.), *Absenteeism* (pp. 194-228). Washington, DC: Jossey-Bass.
- Rusbult, C. E., & Farrell, D. (1983). A longitudinal test of the investment model: The impact on job satisfaction, job commitment, and turnover of variations in rewards, costs, alternatives, and investments. *Journal of Applied Psychology*, 68, 429-438.
- Rush, J. C., Peacock, A. C., & Milkovich, G. T. (1980). Career stages: A partial test of Levinson's model of life/career stages. *Journal of Vocational Behavior*, 16, 347-359.
- Saleh, S. D., & Otis, J. L. (1964). Age and level of job satisfaction. *Personnel Psychology*, 17, 425-430.
- Schein, E. H. (1978). *Career dynamics: Matching individual and organizational needs*. Reading, MA: Addison-Wesley.
- Seashore, S. E. (1974). Job satisfaction as an indicator of the quality of employment. *Social Indicators Research*, 1, 135-168.
- Smith, P. C., Kendall, L. M., & Hulin, C. L. (1969). *The measurement of satisfaction in work and retirement*. Chicago: Rand McNally.
- Super, D. E. (1980). A life-span, life-space approach to career development. *Journal of Vocational Behavior*, 16, 282-298.
- U.S. Department of Labor, Bureau of Labor Statistics. (1986). Household data annual averages. *Employment and Earnings*, 33(1), 154.
- Wanous, J. P. (1980). *Organizational entry: Recruitment, selection, and socialization of newcomers*. Reading, MA: Addison-Wesley.
- Whyte, W. H., Jr. (1957). *The organization man*. Garden City, NY: Doubleday/Anchor Books.
- Wright, J. D., & Hamilton, R. F. (1978). Work satisfaction and age: Some evidence for the "job change" hypothesis. *Social Forces*, 56, 1140-1158.

Received December 5, 1985
 Revision received April 6, 1987
 Accepted March 17, 1987 ■

Kintsch Appointed Editor of *Psychological Review*, 1989-1994

The Publications and Communications Board of the American Psychological Association announces the appointment of Walter Kintsch, University of Colorado, as editor of *Psychological Review* for a 6-year term beginning in 1989. As of January 1, 1988, manuscripts should be directed to

Walter Kintsch
 Department of Psychology
 University of Colorado
 Campus Box 345
 Boulder, Colorado 80309

Manuscript submission patterns for *Psychological Review* make the precise date of completion of the 1988 volume uncertain. The current editor, Martin Hoffman, will receive and consider manuscripts until December 31, 1987. Should the 1988 volume be completed before that date, manuscripts will be redirected to Kintsch for consideration in the 1989 volume.

Comparative Effects of Personal and Situational Influences on Job Outcomes of New Professionals

Stephen M. Colarelli
Central Michigan University

Roger A. Dean
School of Commerce, Economics, and Politics
Washington and Lee University

Constantine Konstans
Fogelman College of Business and Economics, Memphis State University

We investigated the relative and combined effects of personal and situational variables on job outcomes of new professionals. The personal variables were cognitive ability, socioeconomic status, and career goals; the situational variables were job feedback, autonomy, and job context. Data were collected at two times from 280 newly hired, entry-level accountants at "Big Eight" firms. Both personal and situational variables predict job outcomes, but their relative influence depends on the outcome measure. Situational variables account for the most variance in job performance, job satisfaction, and organizational commitment; personal variables account for the most variance in promotability, internal work motivation, and turnover. The findings indicate that job performance does not take care of itself by selecting bright people, but requires constant vigilance and effective systems. The results also suggest that a given result can be achieved through a variety of behavioral science interventions.

Many strategies exist for improving worker productivity and satisfaction. Yet most research focuses on the effectiveness of one strategy at a time. Research on a single approach is important because it adds to the understanding of a strategy's particular dynamics and outcomes, but organizations must address a variety of concerns simultaneously. All organizations must set goals, design jobs, select and train employees, and evaluate performance. Curiously, there is little research that examines the effects of multiple approaches to improving job outcomes. Two recent reviews (Guzzo, Jette, & Katzell, 1985; Locke, Feren, McCaleb, Shaw, & Denny, 1980) examined the comparative effects of behavioral science interventions across a number of studies. Although these reviews illustrate the effectiveness of specific strategies, little could be said about the effects of multiple strategies because most studies examined single interventions (Guzzo et al. 1985, p. 287).

The purpose of this research is to examine the relative and combined influence of personal and situational variables on job outcomes. Our focus on personal *and* situational variables is important for two reasons. First, it furthers our understanding of the comparative influence of personal and situational characteristics on job outcomes. Second, it focuses attention on how

different interventions might be substituted to achieve a given outcome. For example, Allison (1977) found that the same degree of academic achievement in an introductory economics class could be predicted either by selecting students with 200 more points on the Scholastic Aptitude Test or by substituting the best teachers for the worst teachers.

Personal Factors

According to some researchers, variables reflecting characteristics of people are crucial for maximizing job performance. Schmidt, Hunter, and Pearlman (1982), for example, suggested that the probabilities for successful performance are improved when selection and staffing are given priority. Literally hundreds of measures of personal characteristics exist that can be used to select employees. However, several are consistently related to differences in occupational achievement. These are cognitive ability, socioeconomic status (SES), and level of aspiration (career goals).

Cognitive Ability

For years some psychologists have argued that cognitive ability is a critical variable in predicting job and occupational success (Hale, 1982). It is suggested that information-processing and problem-solving skills are important influences on job performance (Schmidt, Hunter, & Pearlman, 1981). Recently, Hunter (1986) argued that cognitive ability affects the acquisition of job knowledge, which in turn influences job performance. A number of industrial psychologists now maintain that cognitive ability tests used for personnel selection could produce huge labor cost savings—as much as "\$16 billion per year for large employers such as the federal government" (Schmidt & Hunter, 1981, p. 1128).

A previous version of this article was presented at the 94th Annual Convention of the American Psychological Association, Washington, DC, 1986.

The authors gratefully acknowledge the suggestions of Terry Beehr, Richard Guzzo, Dan King, Neal Schmitt, Ben Schneider, and John Wainous on earlier versions of this article. The helpful comments of three anonymous reviewers are also appreciated.

Correspondence concerning this article should be addressed to Stephen M. Colarelli, Department of Psychology, Central Michigan University, Mt. Pleasant, Michigan 48859.

Socioeconomic Status

Socioeconomic status and similar background measures are strong predictors of success across a variety of occupations (Jencks, 1979) as well as success within occupations. For example, studies by Dreher, Dougherty, and Whitely (1985) and Pfeffer (1977) suggested that SES is related to managerial success. Family background may be important for career success for several reasons. The role models that parents provide to their offspring have implications for occupational success. Parents of middle and upper SES background may teach the speech and behavior patterns that facilitate career success. They may also inculcate work values respected by employers. And children from more affluent families may have access to experiences that develop effective work habits.

Career Goals

The influence of specific and challenging goals on task performance is well documented (Locke, Shaw, Saari, & Latham, 1981). One would expect, therefore, that an individual with a goal to reach specific career objectives would be more likely to achieve that objective than an employee without such a goal. A long-range goal should foster directed effort over time, strategy development, and receptivity to performance feedback.

Situational Strategies

Others argue that performance (and job attitudes) are more influenced by what happens to people after they are hired. Individual performance is more a function of system qualities and management practices than personal attributes. For example, Roberts, Hulin, and Rousseau (1978, p. 123) argued that the evidence in organizational research suggests that situational characteristics account for a greater portion of responses by people than do personal characteristics. This study examined the influences of three situational factors: autonomy, feedback, and satisfaction with job context.

Autonomy

Autonomy correlates with both performance and satisfaction of professionals (Pelz & Andrews, 1976). It is also a salient concern in the early stages of a professional's career (DeCoster & Rhode, 1971). Autonomy is related to professional productivity for at least three reasons. First, autonomy of workers is congruent with the open-systems principle of *equifinality*—that a system can reach the same end state from varying initial conditions and by a variety of paths. Autonomy allows fuller use of an individual's talents and ingenuity than close supervision and high formalization. Another reason why autonomy relates positively to productivity is that it increases a sense of personal responsibility for getting a job done (Hackman & Oldham, 1980, pp. 79–80). Finally, the absence of autonomy is related to work stress (Hall & Savery, 1986). Close supervision and lack of autonomy create excessive stress that could hinder performance.

Feedback

Feedback is information on how well one is meeting goals. When it is understood, accepted, and acted upon, it enhances

both performance and motivation (Latham & Wexley, 1981). There are several reasons why feedback is important to job performance. Feedback informs individuals about the effectiveness of their performance. It serves a corrective function—showing where and how much improvement is needed. And job feedback also provides goals toward which an individual may strive. In fact, Hogarth (1981) argued that giving periodic feedback during task activities may be a more effective method of assuring adequate performance than estimating the probability of success prior to the beginning of a task. All employees should receive timely feedback. But it is particularly important for new employees because they have had less opportunity to learn whether their performance is on target (Beehr & Love, 1983).

Job Context

Job context includes the broad organizational characteristics in which a job occurs. Although there are a number of factors that may be included under the rubric of job context, the context factors considered here are satisfaction with supervision, coworkers, job security, and compensation. People are unlikely to perform well if they are dissatisfied with the immediate work environment. People who experience a dissatisfying work environment become distracted from their work and focus their energy on coping with unpleasant conditions rather than accomplishing work goals (Oldham, Hackman, & Pearce, 1976).

In sum, this study is concerned with the combined and relative effects of personal and situational variables on the job outcomes of new professionals. The personal variables are cognitive ability, SES, and career goals; the situational variables are job feedback, autonomy, and satisfaction with job context.

Method

Subjects

This study uses data from a larger longitudinal study of newly hired, entry-level accountants. Data were collected from 468 subjects on their first day of work at 11 "Big Eight" accounting firm offices. Each of the Big Eight firms was represented. After one year on the job, subjects and their supervisors were surveyed. Questionnaires were received from 280 subjects (60% response), and information on 395 subjects (84%) was received from supervisors. The analysis sample consisted of 280 subjects (60% of the sample surveyed on their first day at work). The mean age of subjects was 23 years; 61% were men. All offices were located in the southwest United States.

Procedure

During an orientation session, subjects were administered a questionnaire by one of the researchers. The questionnaire included items relating to subjects' background, education, and SES. Immediately following, aptitude tests were administered under standardized conditions.

After one year, a second questionnaire was mailed to the subjects at work. This questionnaire included items about the work environment and measures of job attitudes (these are described in detail later). Subjects were requested to mail their completed questionnaires to a university address, and were assured that individual responses would be anonymous and confidential. At the same time, a performance measurement questionnaire and cover letter were mailed to each subject's supervisor. It requested first-year performance ratings (as per personnel file), and asked additional questions on performance, promotability, and turn-

over. Again, anonymity and confidentiality were assured, and supervisors were asked to mail their completed questionnaires to a university address.

Measures

Cognitive ability. The verbal and reasoning scales of the Ball Aptitude Battery (Layton, 1985) were used to measure cognitive ability. Alternate form reliability for the verbal scale is .98, and test-retest reliability for the reasoning scale is .71. The verbal and reasoning scales were modestly correlated ($r = .25, p < .01$), and were combined into a composite scale. Mean scores on the composite scale could range from 0 to 55.

Undergraduate grade point average (GPA) was also used as a measure of ability. This was considered important because recruiters often use it as an index of ability. This is not totally unfounded; traditional cognitive ability tests do correlate with college grades—although more strongly with freshmen- than senior-year GPAs (Humphreys, 1968)—and have been found to relate to specific areas of job performance (Howard 1986). Grade point average was obtained by self-report on the questionnaire administered to subjects during their first day at work. Self-reports of verifiable biographical data are generally accurate (Cascio, 1975). Scores ranged from 0 to 4.

SES. A measure of socioeconomic status, adapted from Hollingshead and Redlich (1958), was included on the first questionnaire. It was composed of items measuring the occupational and educational status of the subject's father and mother. Occupational status was anchored from 1 (*executives and proprietors of large concerns and major professionals*) to 7 (*unskilled workers*); educational status was anchored from 1 (*graduate training*) to 7 (*less than 7 years of school*). Responses were reverse coded so that higher scores would reflect higher SES. Scores could range from 14 (lowest SES level) to 98 (highest SES level). The highest parental SES score was used to index subjects' SES level. The coefficient alpha for the father's SES scale was .87; and for the mother's SES scale, it was .85.

Career goals. Career goals were measured on the first and second questionnaires. Subjects were asked to check one statement from a list of 11 alternatives that best reflected their long-term career objective. Responses were recoded so that 2 equaled the career goal of becoming a partner in a public accounting firm and all other goals equaled 1. Career goals were dichotomized in this fashion because achieving partnership status is the top of the career ladder in public accounting firms. It is a difficult goal that only a small percentage of employees achieve.

Autonomy, feedback, and satisfaction with job context. Autonomy, feedback, and satisfaction with job context were measured with scales from the Job Diagnostic Survey (JDS; Hackman & Oldham, 1980). Autonomy was measured with the respective scale on the JDS ($\alpha = .74$). Feedback was measured by combining the scales measuring feedback from agents ($\alpha = .82$), feedback from the job itself ($\alpha = .76$), and knowledge of results ($\alpha = .82$). Satisfaction with job context was indexed by combining the JDS context scales of satisfaction with supervision, co-workers, security, and pay ($\alpha = .77, .56, .85$, and $.85$, respectively). These scales were anchored from 1 to 7.

Dependent measures. Job performance was measured by a composite of subjects' annual performance rating and the two following questions: "Would you rehire this person to work for you if he or she were to quit?" (anchored from 1, *definitely not*, to 5, *definitely yes*), and "In general, how easy would it be to find someone who would do as good a job as this person is doing?" (anchored from 1, *very easy*, to 5, *very difficult*). The coefficient alpha for the three items was .82. Promotability was measured by a single item: "How promotable is this person?" anchored from 1 (*definitely not promotable*) to 5 (*has recently been promoted*). General satisfaction ($\alpha = .77$) and internal work motivation ($\alpha = .67$) were measured from the respective scales on the JDS. Organizational commitment ($\alpha = .88$) was measured by Porter, Steers, Mow-

Table 1
Means and Measures of Dispersion

Variable	<i>M</i>	<i>N</i>	<i>SD</i>	<i>V</i>
Personal				
Cognitive ability	42.50	229	6.50	.15
Undergraduate GPA	3.46	279	0.34	.10
Socioeconomic status	80.14	263	15.77	.20
Partnership goal				
First day	1.62	274	0.49	.30
Year one	1.17	280	0.38	.32
Situational				
Autonomy	4.57	280	1.06	.23
Feedback	5.00	280	0.87	.17
Job context	5.10	279	0.77	.15
Dependent				
Performance	3.94	246	0.83	.21
Promotability	4.24	246	0.71	.17
Job satisfaction	4.26	280	1.09	.26
Internal work motivation	5.72	280	0.61	.11
Organizational commitment	5.03	280	0.87	.17
Turnover	1.03	232	0.16	.16

Note. *V* = coefficient of variation; GPA = grade point average.

day, and Boulian's (1974) scale, anchored from 1 to 7. Turnover was assessed by an item on the performance questionnaire sent to supervisors. Supervisors were asked to indicate whether an individual terminated employment with the firm.

Results

Means and Standard Deviations

Table 1 presents the means, standard deviations, and coefficients of variation (*V*). The results show that, as expected, the subjects' average cognitive ability is high, as is their GPA. The average SES of the subjects is also high. The career goal of a majority of the accountants on their first day at work was to become a partner (62%); after one year, the number of subjects who still had this career goal was much lower (17%). The average scores of the situational variables fell above the midpoint. The smallest *V* coefficients among the personal characteristics were cognitive ability scores (.15) and GPA (.10); and the largest variation occurred in year-one partnership goals (.32). The *V* scores on situational variables ranged from .15 on the job context scale to .23 on the autonomy scale. A *V* score represents the ratio of the standard deviation to the mean (Cohen & Cohen, 1975), which provides a common metric for standard deviations. This is useful for comparing the relative sizes of standard deviations based on different raw scores.

Correlations

A correlation matrix of the personal, situational, and dependent variables is presented in Table 2. Correlations in the triangles are correlations within each set of variables. Others are correlations between sets of variables.

Within-set correlations. Most of the personal variables are not intercorrelated, except for moderate correlations between cognitive ability and GPA ($r = .25, p < .01$) and between first-day and year-one partnership goals ($r = .29, p < .01$). All of the

Table 2
Intercorrelation Matrix of All Variables in Study

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Personal														
1. Cognitive ability														
2. College GPA	.25**													
3. Socioeconomic status	.07	.00												
4. Partnership goal, first day	-.06	-.06	-.08											
5. Partnership goal, year one	-.10	-.05	-.06	.29*										
Situational^a														
6. Autonomy	.01	.07	-.01	.02	.13*									
7. Feedback	-.07	.08	.02	.07	.17**	.37**								
8. Job context	.05	.08	.05	.08	.18**	.35**	.44**							
Dependent														
9. Performance	.05	.16*	.06	-.01	.06	.20**	.17**	.31**						
10. Promotability	-.02	.22**	-.04	.09	.04	.11	.00	.19**	.39**					
11. General satisfaction	-.11	.01	-.03	.13*	.29**	.48**	.44**	.53**	.18**	.05				
12. Internal work motivation	-.16*	-.07	-.09	.05	.17**	.16**	.12*	.10	-.05	-.05	.36**			
13. Organizational commitment	-.17**	-.01	.01	.18**	.33**	.31**	.32**	.53**	.14*	.12*	.72**	.41**		
14. Turnover	.05	.04	.14*	-.04	-.08	-.08	-.14*	.00	-.09	-.09	-.11	-.12	-.14*	

Note. Decimal points are omitted. GPA = grade point average.

^a Situational variables were measured after subjects were on the job for one year.

* $p < .05$. ** $p < .01$.

situational variables are significantly intercorrelated. Among the dependent variables, performance is moderately correlated with promotability ($r = .39, p < .01$), and slightly correlated with satisfaction ($r = .18, p < .01$) and organizational commitment ($r = .14, p < .05$). All affective measures are significantly intercorrelated. The high correlation between job satisfaction and organizational commitment ($r = .72, p < .01$) may be because of a close relation between a professional's place of employment and a professional's career development. How one feels about one's organization may influence how one feels about one's job. Turnover is negatively related to organizational commitment ($r = -.14, p < .05$).

Between-set correlations. Among the personal and situational variables there were significant correlations between year-one partnership goals and autonomy ($r = .13, p < .05$), feedback ($r = .17, p < .01$), and job context ($r = .18, p < .01$).

College GPA was the only personal variable that correlated with performance ($r = .16, p < .05$) and promotability ($r = .22, p < .01$). Cognitive ability correlated negatively with internal work motivation ($r = -.16, p < .05$) and organizational commitment ($r = -.17, p < .01$). Socioeconomic status was slightly related to turnover ($r = .14, p < .05$)—the higher the SES, the greater the likelihood of leaving. Partnership goals were significantly related to all affective measures, with the exception that first-day partnership goals were not correlated with internal work motivation.

All three situational variables were positively correlated with performance: autonomy ($r = .20, p < .01$), feedback ($r = .17, p < .01$), and job context ($r = .31, p < .01$). Only job context was related to promotability ($r = .19, p < .01$). All of the situational variables were positively related to the affective dependent measures, and feedback was negatively correlated with turnover ($r = -.14, p < .05$).

Multiple Regression Analyses

Personal and situational variables were entered into hierarchical regression equations. Hierarchical regression accounts for the unique variance of each independent variable (Cohen & Cohen, 1975). The independent variables were entered into the equation based on temporal causal priority. Personal variables were entered first because it was assumed that the subjects possessed personal characteristics before being influenced by the situational variables. The effects of the personal variables are expressed independently of the situational variables, and the variance of the situational variables is expressed with the effects of the personal variables held constant. Variables were also entered in order of temporal priority within sets. Among personal variables, SES was entered first, followed by cognitive ability, GPA, first-day goals, and year-one goals. Among the situational variables, autonomy and feedback were entered prior to job context. There are two reasons for this ordering of situational variables. First, this holds autonomy and feedback constant when job context is entered into the equations. Second, it seems appropriate to enter the more descriptive scales of autonomy and feedback before the more evaluative scale of satisfaction with context. The regression equations are presented in Table 3.

Performance. Personal variables explained 6% of the vari-

Table 3
Relative Influence of Personal and Situational Variables on Job Outcomes: Hierarchical Regression

Independent variable	Performance			Promotability			Job satisfaction			Internal work motivation			Organizational commitment			Turnover		
	β	R^2 cg	R^2 cum	β	R^2 cg	R^2 cum	β	R^2 cg	R^2 cum	β	R^2 cg	R^2 cum	β	R^2 cg	R^2 cum	β	R^2 cg	R^2 cum
Personal		(6)			(11)			(9)			(6)			(15)			(14)	
Socioeconomic status	05	00	00	-05	00	00	-02	00	00	-12	01	01	02	00	00	16	03*	03
Cognitive ability	10	01	01	-04	00	00	-07	00	00	-11	01	03	-14	02	02	02	00	03
College GPA	20	04*	05	25	06**	06	08	01	01	05	00	03	02	00	02	02	00	03
Goal, first day	10	01	06	21	04**	11	16	02	03	13	02	05	21	04**	07	-11	01	04
Goal, year one	-01	00	06	03	00	11	26	06**	10	10	01	06	30	08**	15	-05	00	04
R^2		03	08		00	07			07		02			12				00
Situational		(14)			(8)			(30)			(0)			(25)			(2)	
Autonomy	17	03*	09	17	03*	14	38	14**	24	05	00	06	25	06**	21	04	00	04
Feedback	12	01	10	-06	00	14	23	05**	28	-02	00	06	16	02*	23	-09	01	05
Job context	37	10**	20**	27	05**	19**	38	11**	39**	-02	00	06	48	17**	40**	12	01	06
R^2		16	15		05	36			36		01			37				01

Note. Numbers in parentheses indicate the proportion of variance explained by the respective sets of variables (personal or situational); discrepancies are due to rounding errors. Statistical

ance in performance, with GPA accounting for most of the variance. Situational variables explained 14% of the variance in performance. Job context accounted for the largest share of the variance, followed by autonomy and feedback. Personal and situational variables combined accounted for 20% of the variance in job performance.

Promotability. Personal variables explained the largest portion of the variance in promotability, 11%. College GPA accounted for most of the personal variable variance, followed by first-day partnership goals. Situational variables accounted for 8% of the variance. A total of 19% of the variance in promotability was explained by personal and situational variables combined.

Job satisfaction. The situational variables accounted for 30% of the variance in job satisfaction. They were all strong and significant predictors. The strongest situational predictor was autonomy, followed by job context. Of the variance in job satisfaction, 9% was accounted for by personal variables, with year-one partnership goals accounting for most of this variance. The combined situational and personal variables accounted for 39% of the variance in job satisfaction.

Internal work motivation. Neither the personal nor situational variables accounted for a significant portion of the variance in internal work motivation.

Organizational commitment. Situational variables accounted for 25% of the variance in organizational commitment. Job context accounted for most of this, followed by autonomy. Personal variables accounted for 15% of the variance, most of which was attributable to first-day and year-one partnership goals. Both situational and personal variables accounted for 40% of the variance in organizational commitment.

Turnover. Socioeconomic status was the only variable that accounted for a significant amount of the variance in turnover (3%).

The shrunken R^2 s (\tilde{R}^2) did not differ appreciably from those cumulative R^2 s that reached statistical significance. This suggests that the regression results are fairly robust.

Job Context

It is appropriate to examine the job context measure in more detail for two reasons. First, it had the largest beta weight of the situational variables in four out of the six outcome measures. Second, it is composed of four subscales. Table 4 presents the correlations between the four job context subscales and the six outcome measures. The context subscales of satisfaction with supervision, coworkers, and job security are all significantly related to job performance. Job security correlated more strongly with performance than any of the other context subscales. Job security was the only context scale that correlated with promotability. All context subscales were significantly correlated with general job satisfaction, and a similar pattern of correlations occurred with organizational commitment. Supervision and coworkers correlated modestly with internal work motivation. None of the context variables correlated with turnover.

Discussion

This research examined the influence of personal and situational variables on job outcomes of new professionals. The re-

Table 4
Intercorrelation Matrix of Subscales of Job Context Scale and Dependent Variables

Variable	1	2	3	4	5	6	7	8	9	10
1. Supervision	—									
2. Coworkers	.44** (.30**)	—								
3. Security	.39** (.30**)	.20** (.09)	—							
4. Pay	.29** (.17**)	.08 (-.05)	.32** (.26**)	—						
5. Performance	.26** (.20**)	.17** (.11)	.35** (.31**)	.06 (.01)	—					
6. Promotability	.14* (.13*)	.09 (.07)	.27** (.27**)	.00 (-.01)	.39** (.38**)	—				
7. Satisfaction	.49**	.41**	.29**	.29**	.18**	.05	—			
8. Internal work motivation	.13* (-.05)	.25** (.11)	-.03 (-.15*)	.01 (-.11)	-.06 (-.13*)	-.05 (-.08)	.36**	—		
9. Organizational commitment	.45** (.15**)	.47** (.28**)	.30** (.14*)	.29** (.12*)	.14* (.03)	.12* (.12)	.72**	.41** (.23**)	—	
10. Turnover	-.08 (-.03)	-.06 (-.02)	.04 (.07)	.07 (.10)	.00 (.02)	-.09 (-.09)	-.11	-.12 (-.09)	-.14* (-.09)	—

Note. Due to missing data, *N*s range from 232 to 280. Decimal points are omitted. Numbers in parentheses are partial correlations, holding job satisfaction constant.

* $p \leq .05$. ** $p \leq .01$.

sults indicated that both personal and situational variables predict job outcomes, but their relative influence depends on the outcome measure. Situational variables accounted for the most variance in job performance, general job satisfaction, and organizational commitment. Personal variables accounted for the most variance in promotability, internal work motivation, and turnover. The combined set of variables explained the most variance in organizational commitment and job satisfaction, somewhat less in job performance and promotability, and relatively little in turnover and internal work motivation.

Impact of Personal and Situational Variables

Performance and promotability. A telling finding was that the largest proportion of the variance in performance was explained by situational variables, whereas the largest proportion of the variance in promotability was explained by personal variables. This suggests that major influences on job performance are job conditions and competent management. Promotions, on the other hand, may have less to do with performance effectiveness than individual adaptation or management stereotypes. Such factors as anticipated performance as a partner, ability to fit in, and supervisor expectations may contribute to promotability above and beyond performance. A Pygmalion effect may also account for some of the variance in performance and promotability ratings if supervisors were aware of background information on employees, such as college GPA.

Increasing job satisfaction. Several scholars have recently suggested that personal variables may be more potent causes of job satisfaction than previously believed, and that selection strategies could be used to increase job satisfaction (Staw & Ross, 1985). The results of this study show that personal variables explain a smaller percentage of the variance in job satis-

faction than do situational variables. As an approach to increasing job satisfaction, selecting happy people may not always be practical when more job satisfaction can be produced by situational strategies. One might argue that an emphasis on personal qualities as causes of satisfaction (or performance) reflects the fundamental attribution error—overestimating the role of personal factors and underestimating the role of situational factors in influencing behavior (Ross, 1977).

Limitations

The variance in cognitive ability and GPA was restricted. This is a limitation in the sense that the size of correlations is influenced by the size of the variance on predictor and criterion variables. However, the restriction in range in cognitive ability and GPA in the present data reflects the variance in ability of the applicant pool from which major professional firms select new employees. Moreover, the coefficient of variation for cognitive ability in the present sample was the same as the coefficient of variation for job context and only slightly smaller than the coefficient of variation for feedback, both of which had stronger correlations with outcome variables than did cognitive ability.

One might ask to what extent do the context subscales assess the "objective" characteristics of the job context? In other words, do the context factors exist independently of subjects' feelings of satisfaction toward them? The data are less clear on this issue. Table 4 shows the correlations between the four context scales and outcome measures with general satisfaction partialled out. Although the partials are generally smaller than the whole correlations, the *pattern* of correlations is similar, and many of the partial correlations are significant. Furthermore, the pattern of correlations between general satisfaction and job outcomes differs from the pattern of context satisfaction and

job outcomes (see Table 2). This suggests that the context scales measure objective job characteristics as well as satisfactions. Caution is also warranted because the context variables and some of the dependent variables were measured on the same instrument, and therefore share common method variance.

An important issue in this study was the choice of variables. It could be argued that other variables are more representative of the classes of personal and situational variables. To be sure, we cannot claim to have selected a definitive set of personal and situational variables, although the variables included in the present study are widely cited in the literature. Continued multivariate research with these and other variables is strongly encouraged.

In interpreting comparisons of multiple variables, caution must be exercised. The appropriateness of comparisons depends on the equivalence in distributions and on the equivalence in the fidelity of measurement (Cooper & Richardson, 1986). Our interpretation of the results is compatible with the distributions of the variables. However, with differences in the fidelity of measurement (e.g., reliabilities), interpretation is more problematic. Therefore, the results should be interpreted with care. The comparisons should be approached as trends rather than as precise differences.

Finally, the sample is both a strength and a limitation. It is a strength in that it is a large, representative sample of one particular group of employees—first-year accountants in Big Eight firms. As such, the findings should have strong external validity when applied to major accounting firms. The findings should also be applicable to professional employees in similar professional bureaucracies. On the other hand, the sample is not representative of the work force in general. As a result, conclusions about nonprofessional employees in less selective organizations should be made cautiously.

Implications

Personnel selection. There are at least two implications of these results for the use of cognitive ability tests for personnel selection. First, range restriction is a fact of organizational life in some situations. For example, the range of applicants may be restricted by minimum requirements or self-selection. People tend to apply to organizations or for positions that are consistent with their self-image and perceived level of competence (cf. Korman, 1977; Tom, 1971). Thus, the design and estimated benefits of selection programs should reflect the normally occurring distributions on predictors and criteria in a given organization. Second, the results of this study, and others (Hough, 1984; Kraut, 1969), suggest that the predictive power of cognitive ability tests is modest in organizations that attract talented applicants. Beyond a specific cognitive threshold, other factors may be more influential determinants of performance.

The results also suggest that generalities about SES as a predictor of successful work behavior may be inappropriate. In the present study, SES was unrelated to performance or any of the affective measures. Socioeconomic status is a proxy for certain attitude and behavior constellations. Work values and habits may be more a function of specific child-rearing practices, adult role models, and values in the home than is SES per se. On the other hand, those SES characteristics assumed to be associated

with career advancement may be less relevant at beginning levels in public accounting, where technical proficiency is emphasized. They may be more associated with performance and promotability at higher levels where social, communication, and leadership skills are more important.

System characteristics. The results of this study suggest that, in the context of Big Eight accounting firms, system characteristics have a stronger influence on individual performance than do personal characteristics. Given adequate talent, an environment that fosters the development and utilization of human abilities is critical to effective job performance. If the work environment is not effectively managed, moderate variation in personal characteristics will have little influence on job performance and satisfaction (cf. Fiedler, Potter, Zais, & Knowlton, 1979).

When managers emphasize personal characteristics as causes of performance, they may be doing so to absolve themselves (management) of responsibility for performance. Research suggests that more experienced and competent supervisors recognize the importance of situational characteristics in influencing performance, whereas less experienced managers tend to make attributions to personal characteristics (Mitchell & Kalb, 1982). More effective supervisors also tend to recommend changes in the environment to improve performance (Klemp, Munger, & Spencer, 1977).

An unwarranted emphasis on personal characteristics can provide a spurious justification for not putting *continuous* effort into human resource programs. If it is assumed that performance takes care of itself once the right people are selected, then managers may not see any merit in working to improve an individual's performance. But performance does not take care of itself by selecting bright people (Fiedler et al., 1979; Schneider & Schmitt, 1986, pp. 408–409). Good performance requires constant vigilance and effort (cf. Hogarth, 1981).

Areas for Future Research

Production functions. Behavioral science has advanced past the point of questioning “whether or not” an intervention is valid to asking “how large” is its effect and “under what circumstances” will it work (Guzzo et al., 1985). An important area for future research, therefore, would be developing models of the contributions of *multiple* behavioral science interventions to job outcomes. The concept of the *production function* from the economic literature is relevant. According to Samuelson (1976),

The production function is the technical relationship telling the maximum amount of output capable of being produced by each and every set of specified inputs (or factors of production). It is defined for a given state of technical knowledge. (p. 537)

The results of this study suggest that organizational psychologists can conceptualize job performance and job satisfaction as outputs resulting from “production” inputs. It now seems appropriate to utilize the *existing state of behavioral science technology* to develop production functions for different outcomes. This is important in a postindustrial economy when productivity is increasingly a function of human resources.

Equifinality and substitutability. *Equifinality* is a principle

of open-systems theory that states that "a system can reach the same final state from differing initial conditions and by a variety of paths" (Katz & Kahn, 1966, pp. 25–26). The present study demonstrates the viability of that principle with behavioral science interventions. A related concept from the economic literature is *substitutability*: the degree to which two or more inputs can be substituted for each other to produce a specific amount of an output (Samuelson, 1976). Substitutability implies a procedure for formalizing multiple causal mechanisms. Incorporating the concept of substitutability into the practice of industrial–organizational psychology would be advantageous. It implies flexibility in using interventions, and changes the focus of industrial–organizational psychology from techniques to results.

For example, consider the plausibility of substituting a selection strategy for a job context strategy with accountants. The beta weights in Table 3 provide an index of the substitutability of the variables used in the present study. An accounting firm would have to select applicants with college GPAs almost twice as high as those of the present applicants to achieve equivalent results in job performance. If firms were to use cognitive ability tests in their selection process, they would have to select applicants with test scores 3.7 times as high as those they are currently selecting to have the same effect on job performance as job context. Here, an emphasis on selection would have little effect on performance, whereas situational strategies would have the greatest payoff. On the other hand, if there were a large variance in ability among applicants, then a selection strategy might be an effective method of increasing productivity.

Management values and interventions. Finally, it is important to examine the relation between behavioral science strategies and management values. As this research suggests, managers can choose from a variety of behavioral science options to achieve a desired outcome. It would be fruitful to examine how management values and assumptions about human nature influence the interventions that organizations use, or are inclined to use. How managers think about causality may be as relevant to the use of behavioral science interventions as the actual relation between interventions and behavior.

References

- Allison, E. (1977). *Educational production function for an introductory economics course* (Discussion paper No. 545). Cambridge, MA: Harvard University, Harvard Institute of Economic Research.
- Beehr, T. A., & Love, K. G. (1983). A meta-model of the effects of goal characteristics, feedback, and role characteristics in human organizations. *Human Relations*, 36, 151–166.
- Cascio, W. F. (1975). Accuracy of verifiable biographical information blank responses. *Journal of Applied Psychology*, 60, 767–769.
- Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Cooper, W. H., & Richardson, A. J. (1986). Unfair comparisons. *Journal of Applied Psychology*, 71, 179–184.
- DeCoster, D. T., & Rhode, J. G. (1971). The accountant's stereotype: Real or imagined, deserved or unwarranted. *The Accounting Review*, 46, 651–664.
- Dreher, G. F., Dougherty, T. W., & Whitely, B. (1985). Generalizability of MBA degree and socioeconomic effects on business school graduates' salaries. *Journal of Applied Psychology*, 70, 769–773.
- Fiedler, F. E., Potter, E. H., III, Zais, M. M., & Knowlton, W. A., Jr. (1979). Organizational stress and the use and misuse of managerial intelligence and experience. *Journal of Applied Psychology*, 64, 635–647.
- Guzzo, R. A., Jette, R. D., & Katzell, R. A. (1985). The effects of psychologically based intervention programs on worker productivity: A meta-analysis. *Personnel Psychology*, 38, 275–291.
- Hackman, J. R., & Oldham, G. R. (1980). *Work redesign*. Reading, MA: Addison-Wesley.
- Hale, M. (1982). History of employment testing. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies, Part 2: Documentation section* (pp. 3–38). Washington, DC: National Academy Press.
- Hall, K., & Savery, L. K. (1986). Tight rein, more stress. *Harvard Business Review*, 64, 160–164.
- Hogarth, R. M. (1981). Beyond discrete biases: Functional and dysfunctional aspects of judgemental heuristics. *Psychological Bulletin*, 90, 197–217.
- Hollingshead, A. B., & Redlich, F. C. (1958). *Social class and mental illness: A community study*. New York: Wiley.
- Hough, L. M. (1984). Development and evaluation of the "Accomplishment Record" method of selecting and promoting professionals. *Journal of Applied Psychology*, 69, 135–146.
- Howard, A. (1986). College experiences and managerial performance. *Journal of Applied Psychology*, 71, 530–552.
- Humphreys, L. G. (1968). The fleeting nature of the prediction of college academic success. *Journal of Educational Psychology*, 59, 375–380.
- Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior*, 29, 340–362.
- Jencks, C. (1979). *Who gets ahead?* New York: Basic Books.
- Katz, D., & Kahn, R. L. (1966). *The social psychology of organizations*. New York: Wiley.
- Klemp, G. O., Munger, M. T., & Spencer, L. M. (1977). *Analysis of leadership and management competencies of commissioned and non-commissioned Naval officers in the Pacific and Atlantic fleets* (Final report). Boston: McBer & Company.
- Korman, A. K. (1977). *Organizational behavior*. Englewood Cliffs, NJ: Prentice-Hall.
- Kraut, A. I. (1969). Intellectual ability and promotional success among high level managers. *Personnel Psychology*, 22, 281–290.
- Latham, G. P., & Wexley, K. N. (1981). *Increasing productivity through performance appraisal*. Reading, MA: Addison-Wesley.
- Layton, W. L. (1985). Review of the Ball Aptitude Battery. In J. V. Mitchell, Jr. (Ed.), *The ninth mental measurements yearbook* (Vol. 1, pp. 123–125). Lincoln: University of Nebraska, Buros Institute of Mental Measurements.
- Locke, E. A., Feren, D. B., McCaleb, V. M., Shaw, K. N., & Denny, A. T. (1980). The relative effectiveness of four methods of motivating employee performance. In K. D. Duncan, M. M. Gruneberg, & D. Wallis (Eds.), *Changes in working life* (pp. 363–388). New York: Wiley.
- Locke, E. A., Shaw, K. N., Saari, L. M., & Latham, G. P. (1981). Goal setting and task performance: 1969–1980. *Psychological Bulletin*, 90, 125–152.
- Mitchell, T. R., & Kalb, L. S. (1982). Effects of job experience on supervisor attributions for a subordinate's poor performance. *Journal of Applied Psychology*, 67, 181–188.
- Oldham, G. R., Hackman, J. R., & Pearce, J. L. (1976). Conditions under which employees respond positively to enriched work. *Journal of Applied Psychology*, 61, 395–403.
- Pelz, D. C. & Andrews, F. M. (1976). *Scientists in organizations* (Rev. ed.). Ann Arbor: University of Michigan, Institute for Social Research.

- Pfeffer, J. (1977). Effects of an MBA and socioeconomic origins on business school graduates' salaries. *Journal of Applied Psychology*, 62, 698-705.
- Porter, L. W., Steers, R. M., Mowday, R. T., & Boulian, P. V. (1974). Organizational commitment, job satisfaction, and turnover among psychiatric technicians. *Journal of Applied Psychology*, 59, 603-609.
- Roberts, K. H., Hulin, C. L., & Rosseau, D. M. (1978). *Developing an interdisciplinary science of organizations*. San Francisco: Jossey-Bass.
- Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10, pp. 173-220). New York: Academic Press.
- Samuelson, P. A. (1976). *Economics* (10th ed.). New York: McGraw-Hill.
- Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist*, 36, 1128-1137.
- Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: A red herring. *Journal of Applied Psychology*, 66, 166-185.
- Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1982). Assessing the economic impact of personnel programs on work force productivity. *Personnel Psychology*, 35, 333-347.
- Schneider, B., & Schmitt, N. (1986). *Staffing organizations* (2nd ed.). Glenview, IL: Scott, Foresman.
- Staw, B. M., & Ross, J. (1985). Stability in the midst of change: A dispositional approach to job attitudes. *Journal of Applied Psychology*, 70, 469-480.
- Tom, V. R. (1971). The role of personality and organizational images in the recruiting process. *Organizational Behavior and Human Performance*, 6, 573-592.

Received November 7, 1986

Revision received May 20, 1987

Accepted April 22, 1987 ■

Delworth Appointed Editor of *Professional Psychology: Research and Practice*, 1989-1994

The Publications and Communications Board of the American Psychological Association announces the appointment of Ursula M. Delworth, University of Iowa, as editor of *Professional Psychology: Research and Practice* for a 6-year term beginning in 1989. As of January 1, 1988, manuscripts should be directed to

Ursula Delworth
University of Iowa
College of Education
360 LC
Iowa City, Iowa 52242

Manuscript submission patterns for *Professional Psychology: Research and Practice* make the precise date of completion of the 1988 volume uncertain. The current editor, Norman Abeles, will receive and consider manuscripts until December 31, 1987. Should the 1988 volume be completed before that date, manuscripts will be redirected to Delworth for consideration in the 1989 volume.

Effect of Rater Training on Rater Accuracy: Levels-of-Processing Theory and Social Facilitation Theory Perspectives

Timothy R. Athey
Colorado State University

Robert M. McIntyre
Old Dominion University

We use levels-of-processing theory and social facilitation theory to explain the effect of training format and group size on distance and correlation accuracy, leniency-severity, halo, retention of training and pretraining information, and subject arousal. The training factor included frame-of-reference (FOR) training, information only (INFO) training, and no training (NOT). Group size was $n = 1$, $n = 6$, and $n = 12$, respectively. A total of 108 subjects, randomly assigned to one of nine Training \times Group Size conditions, viewed and rated videotaped lectures. Results indicated that FOR training effected improved retention of training information, improved distance accuracy, and less halo over INFO training or no training ($p < .05$). Group size significantly affected the retention of pretraining information but not the retention of training information. Discussion centers on the components of FOR training responsible for improved rating accuracy and error, the relation between rating knowledge and rating accuracy, and implications for future research.

Although considerable progress has been made recently in improving the effectiveness of rater training programs, little is yet understood about the cognitive operations involved in the rater training process. Given the current emphasis in the field on discovering the cognitive processes underlying performance evaluation in general (Cooper, 1981; Feldman, 1981; Landy & Farr, 1980; Nathan & Lord, 1983), it is surprising that this same approach has not been applied to the rater training process. Rater training research essentially represents a laboratory experience designed to alter individual information-processing characteristics by teaching raters to categorize behavioral observations into discrete performance dimensions. Thus, the rater training research paradigm provides a mechanism for identifying the basic cognitive operations influenced by rater training and related to improved rating accuracy.

Frame-of-Reference Training

Current research in rater training demonstrates a change in focus from error avoidance training to more proactive rater accuracy approaches. One important training method in this genre is that of Bernardin and Buckley (1981) called *frame-of-reference training*. This approach emphasizes the establishment of a common frame of reference from which raters can evaluate ratee performance. More recently, McIntyre, Smith, and Hassett (1984) have described frame-of-reference training as possessing the following components: (a) a description of the work to be evaluated, (b) practice and feedback with ratings, and (c) behavioral rationales for ratings provided by the expert raters. The intended effect of this training process is to facilitate more

accurate ratings by standardizing raters' perceptions of the behaviors associated with good and poor performance. Although McIntyre et al. (1984) demonstrated that frame-of-reference training was superior to traditional error training in improving rating accuracy, the cognitive changes occurring as a result of the training experience itself remain unclear. The purpose of the present study is to investigate the fundamental cognitive outcomes resulting from frame-of-reference training and, also, to see how these outcomes may be related to improved rating accuracy.

Bloom's Taxonomy of Educational Objectives

Rater training is viewed as a specific example of a group instructional situation. Consequently, the rater training process should be governed by the same fundamental learning principles applied to other instructional paradigms. A general model for the instructional process is provided in Bloom's taxonomy of educational objectives (Bloom, Engelhart, Walker, Furst, & Krathwohl, 1956). The Bloom model represents the instructional process as a hierarchy of cognitive operations ranging from simple knowledge of instructional content at the most fundamental level to the evaluation of complex stimuli using that content at the highest level. Bloom's taxonomy implies that the capacity of raters to effectively evaluate complex stimulus material, such as videotaped performances, depends on the degree to which they have actually comprehended the information provided in a given rater training program. What is needed is additional insight into the processes that account for the learning of rater training content and the role this learning plays in improving rater accuracy. Levels-of-processing theory and social facilitation theory help provide this insight.

Levels-of-Processing Theory

Levels-of-processing theory (Craik & Lockhart, 1972) provides a basis for understanding how different rater training for-

We express our gratitude to Kevin R. Murphy and his colleagues for providing us with videotaped stimuli.

Correspondence concerning this article should be addressed to Timothy R. Athey, Department of Psychology, Colorado State University, Fort Collins, Colorado 80523.

mats may affect the learning of training content. The theory represents a significant departure from traditional "stage theories" of memory (Atkinson & Shiffrin, 1968) in that it does not hypothesize the existence of internal structures, such as short-term and long-term memory, to explain the memory process. Rather, the theory defines memory as a function of the depth of perceptual processing required by any given information. Proponents of the theory (Craik & Lockhart, 1972; Jacoby & Craik, 1979) argued that deeper levels of perceptual processing involving greater cognitive elaboration will result in a more distinctive memory trace and better retention of the information. Subsequent research on the levels-of-processing model has, in general, confirmed the fundamental principles originally formulated by Craik and Lockhart (Moeser, 1983; Reed, 1982; Reynolds & Flagg, 1983).

Rater accuracy training essentially involves the clarification of connections between performance dimensions and actual behavior. Trainees must develop cognitive associations between dimensions and behaviors by essentially rote-memorizing specific dimension-behavior pairings. The level at which rater training material is processed may explain the relative success of different training formats. If raters engage in the dimension-behavior pairing process at a superficial level, their retention of training content and the accuracy of their ratings will be less than optimal.

The effect of frame-of-reference training is to enhance the pairing process by engaging raters in a relatively deep level of cognitive processing of the training information through the use of behavioral examples and feedback of expert ratings. Levels-of-processing theory suggests that frame-of-reference training improves accuracy of rating by improving the retention of training content. Bernardin, Abbott, and Cooper (1985) offered preliminary support for this hypothesis in reporting a significant relation between learning and accuracy.

Social Facilitation Theory

Because rater training typically takes place in groups, it is appropriate to consider research on the effect of group size. Social facilitation theory hypothesizes that memory is affected by the size of the group within which training occurs. The theory (Zajonc, 1965, 1980) states that the presence of others in a learning or performance situation increases arousal level and, in turn, affects performance. Although much of the research on social facilitation has involved motor tasks, there is evidence that the presence of others also affects the comprehension and retention of information (Beatty, 1980; Klausmeier, Wiersma, & Harris, 1963). McIntyre and Athey (1985), in a previous study concerned with the effect of group size on learning of training material and accuracy of rating, found that small groups (consisting of 3, 4, or 5 trainees) comprehended significantly more pretraining information than did larger groups (consisting of 15, 16, 17, or 18 trainees). In addition, group size marginally affected comprehension of training content. Because rater training group size can vary from study to study, it is important to investigate its effect on the retention of training information and accuracy of ratings.

The purpose of this study was, therefore, threefold: (a) to confirm previous findings (McIntyre et al., 1984) that frame-of-

reference training does improve rating accuracy, (b) to investigate the effect of frame-of-reference training on retention of training content and subsequent improvements in rating accuracy, and (c) to follow up previous research on how training group size may moderate the training effect on retention and accuracy. The specific hypotheses of interest in this study were as follows:

1. A frame-of-reference (FOR) group will demonstrate significantly better retention of training information than a group receiving only the verbal specification of performance dimensions (INFO). Both groups will demonstrate significantly better retention of training information than a group receiving no information at all (NOT).

2. The FOR group will demonstrate significantly more accurate ratings than the INFO group. Both groups will demonstrate significantly more accurate ratings than the NOT group.

3. Group size will affect the retention of training information within the FOR and INFO groups such that the "groups" consisting of one individual will retain more than the 6-member groups, and both will retain more than the 12-member groups. To the extent that group size is found to affect retention of training information, a similar effect will occur for rating accuracy.

4. Group size will affect the retention of pretraining orientation information such that the 1-member "groups" will retain more than the 6-member groups, and both will retain more than the 12-member groups. In addition, self-report measures of arousal should increase with group size.

Method

Overview of the Procedure

Subjects were placed in one of three training conditions and one of three group sizes. They viewed one videotaped minilecture during the training period. Finally, they viewed three videotaped minilectures and rated the quality of each lecturer's performance.

Subjects

In all, 108 undergraduate students (60 women, 48 men) participated in the study as a part of the research requirements for an introductory psychology course at Colorado State University during the fall of 1983.

Stimulus Material

Four videotaped lectures developed by Murphy, Garcia, Kerkar, Martin, and Balzer (1982) were used as the rating stimuli. Two of the lectures were about self-fulfilling prophecy and two were about crowding and stress. Each lecture was approximately 6 min in length, and all of the videotapes were presented in the same order to all of the subjects. The lectures were performed by male drama students and were carefully scripted by Murphy et al. to ensure realistic performances.

Rating Scale

A 12-item rating scale developed by McIntyre et al. (1984) was used as the measure of rating response to the videotapes. The scale consists of items associated with four basic factors of teaching effectiveness identified by Costin (1974): organization, clarity of communication, elocutionary skills, and intellectual stimulation. Scale items were presented in a 7-point Likert-type format with responses ranging from *strongly agree* to *strongly disagree*. The following two items exemplify the scale:

"He followed a logical sequence of thought in his lecture," and "he provided relevant answers to questions." Coefficient alpha for the scale was found by McIntyre et al. (1984) to be .87.

Target Scores

Target scores, a term suggested by Dickinson (personal communication; summer 1985) as an appropriate replacement for "true scores", for each of the videotapes were those used by McIntyre et al. (1984) and were based on ratings of the lectures provided by a group of graduate students in the industrial-organizational psychology program at Colorado State University. These students qualified as expert raters on the basis of their experience as students and lecturers and their knowledge of the field of performance appraisal. Two groups of three students independently viewed and rated the lectures with the rating scale used in this study. Target scores were obtained by computing the mean of the consensual ratings of the two groups on the 12 items. Interrater agreement was computed by correlating the consensual ratings of the two groups on the 12 items for each tape and then computing the median of the resulting four correlations. The interrater agreement value was .641. McIntyre et al. (1984) explained that this rather modest value was probably due to restriction of range of the expert ratings across the 12 items for each tape and possibly to the lack of clarity of the lecturer performance domain.

Rater Training

Three training formats were used in this study. They are referred to as no training (NOT), information-only training (INFO), and frame-of-reference training (FOR).

Members of the NOT group received a general introduction to the experiment, a brief explanation of the rating scale used in the study, and one practice trial involving the viewing and rating of a videotaped lecture. Care was taken not to provide subjects in this training condition with cues regarding the behaviors corresponding to performance dimensions. Following the practice rating exercise, subjects viewed and rated the remaining three test videotapes and completed all questionnaires.

Members of the INFO group received all information and practice provided to the NOT groups. In addition, subjects in the INFO group received visual and oral presentations of the performance items and the behavioral components corresponding to each of the 12 rating scale items. This was accomplished by providing subjects with a photocopied handout listing each rating scale item and the behaviors corresponding to performance on each item. Subjects were instructed to read the handout while the investigators orally reviewed and discussed the information on the handout. To ensure that the subjects had to rely on memory of the training content when rating the subsequent videotapes, subjects were not allowed to take notes on the training content. All of the handouts were collected prior to presenting the videotapes to be rated. Following the INFO training, subjects rated the three test videotapes and completed all questionnaires in the same order as the NOT group. The INFO training lasted approximately 20 min.

The FOR training used in this study is the same as that used by McIntyre et al. (1984). Members of this group received all information and practice provided in the INFO and NOT groups. When reviewing the photocopied handout of performance dimensions used in the INFO group, however, subjects in the FOR group also received evaluative and behavioral cues corresponding to each rating scale item. Subjects were provided with target scores (expert ratings) on each of the rating scale items for the practice videotape and a replay of the portion of the videotape exemplifying performance on each item. Thus, subjects received not only the informational content of the training provided to those in the INFO group, but also feedback on the accuracy of their practice

ratings (evaluative cues) and behavioral examples (behavioral cues) corresponding to those ratings. As in the INFO group, subjects were not allowed to take notes on the training content. All handouts were collected prior to presenting the remaining three videotapes. Following FOR training, subjects completed all questionnaires in the same order as the NOT and INFO groups. The FOR training lasted approximately 30 min.

Dependent Variables

The seven dependent variables used in this study are as follows:

Retention of training information. We used a 12-item test, developed in a previous study by McIntyre and Athey (1985), to measure the subjects' retention of training information. These 12 items required subjects to match each rating scale item with its corresponding behavioral component(s), thus providing a direct measure of subjects' retention of the training content. This scale was administered to the NOT groups as well as to the training groups in order to provide an estimate of the baseline knowledge of performance dimensions possessed by untrained subjects.

Retention of pretraining information. Five additional short-answer questions were included to assess subjects' retention of pretraining information provided in the initial experimental orientation that all of the groups received. These items were included to assess the effect of group size on learning of general experimental information provided prior to training. The items required subjects to recall specific statements made by the experimenter regarding the stated purpose of the study, the nature of the videotapes, and instructions on the use of the rating scale.

Rating accuracy. Correlation and distance accuracy indices described in McIntyre et al. (1984) were used in this study. Correlation accuracy was computed in the following way. For each ratee, the profile of 12 ratings was correlated with the mean profile of expert ratings. The mean of the three *z*-transformed correlations yielded a summary accuracy index similar to Borman's (1975), except that correlations were computed within profiles for each ratee. This metric has been discussed by Cronbach (1955) as a "*Q*-correlation" differential accuracy measure.

Distance accuracy was calculated by determining the average non-squared euclidean distance of each subject's profiles of ratings of the three ratees from the corresponding expert profiles. Correlation accuracy reflects the parallelism between subjects' ratings and target scores. Distance accuracy reflects the level difference between them.

Rating error. In addition to rating accuracy, measures of halo and leniency-severity were used (McIntyre et al., 1984). *Halo* was defined as the mean difference between the variance of obtained ratings per ratee and the variance of expert ratings for that ratee, computed across the three ratees. A positive halo score reflects less rating variability (i.e., halo) than should exist.

Leniency-severity was defined as a rater's tendency to assign ratings that were higher (more lenient) or lower (more severe) than expert ratings. The measure of leniency-severity used in this study reflects the mean signed difference of obtained ratings from target scores computed across all three rating scales. A negative value on this measure reflects leniency, whereas a positive value reflects severity.

Arousal. The 11-item Test Anxiety Profile (TAP; Oetting & Deffenbacher, 1980) was used as a measure of arousal on two dimensions: feelings of anxiety (FA), and thought interference (TI). The FA scale measures an individual's perceived physiological-emotional arousal level. The TI scale reflects the degree to which the individual's thought processes are disrupted by perceived physiological-emotional arousal. Coefficient alpha for the TAP ranges from .88 to .96 for the FA scale and from .90 to .96 for the TI scale. Although the TAP has traditionally been used as a measure of test anxiety, its validity as a more general measure

Table 1
Means and Standard Deviations of Dependent Variables by Training Group

Variable	Group					
	FOR		INFO		NOT	
	M	SD	M	SD	M	SD
Feelings of anxiety	17.61	4.51	15.33	4.88	14.58	6.04
Thought interference	19.31	5.20	18.94	6.57	17.56	6.16
Training retention	7.58	3.14	6.38	2.80	4.60	2.10
Orientation retention	3.08	1.25	3.03	1.28	3.14	1.22
Distance accuracy	14.44	3.32	17.48	4.48	17.35	4.61
Correlation accuracy	.51	.19	.54	.19	.49	.20
Leniency/severity	8.00	6.00	9.43	8.88	7.84	8.13
Halo	1.54	0.67	2.08	0.79	2.07	0.80

Note. FOR = frame of reference; INFO = information only; NOT = no training.

of performance-related state anxiety is supported by Oetting and Deffenbacher (1980). The following exemplify items on the scale:

How I feel now		
Loose	____: ____: ____: ____: ____: ____:	Tight
Helpless	____: ____: ____: ____: ____: ____:	Secure
What my thoughts are like now		
Unsure	____: ____: ____: ____: ____: ____:	Sure
Clear	____: ____: ____: ____: ____: ____:	Confused

Results

A 3 × 3 (Training × Group Size) analysis of variance (ANOVA) was conducted with each of the eight dependent variables. Each of the nine cells in the experimental design contained exactly 12 subjects. Means and standard deviations for all dependent variables by training group are presented in Table 1.

Training Effects

Retention. As hypothesized, training had a significant effect on the retention of rater training information, $F(2, 99) = 7.78$, $p < .001$, $sR^2 = .14$. Newman-Keuls analyses revealed that, as predicted, the FOR group accurately remembered more of the training content than did the INFO group. In addition, the FOR and INFO groups knew more than did the NOT group, but we found no effect for training on the retention of pretraining information. These results support the hypothesis that FOR training improves retention of rater training information.

Rating accuracy. Training significantly affected distance accuracy, $F(2, 99) = 6.06$, $p < .01$, $sR^2 = .11$. Newman-Keuls analyses revealed that, as predicted, the FOR group provided significantly more accurate ratings than did the INFO or NOT groups. No difference in distance accuracy was found between the INFO and NOT groups, however. No significant effect for training on correlation accuracy was observed.

The reliability of the rating accuracy measures was further investigated in an effort to explain the lack of significant training effect for correlation accuracy. The approach used for this analysis represents the reliability of the accuracy measures within raters, across each of the three ratees (i.e., $N = 108$, $k = 3$), and reflects what has been referred to as *intraclass correlation*

(Borman, 1975; Snedecor, 1946). The results of this analysis revealed the reliabilities of the correlation accuracy and distance accuracy measures to be .21 and .26, respectively. Further investigation of the mean values of each accuracy measure across the three ratees revealed correlation accuracy values of .75, .18, and .61, and distance accuracy values of 12.4, 19.5, and 18.1, respectively. As the values indicate, accuracy for both measures was lower for the second of the three ratees, an effect observed within all three treatment groups. This phenomenon may have contributed to the low reliability for both accuracy measures. Review of the second videotaped performance, along with the expert ratings for this performance, revealed no obvious reason for this observed effect. Because interrater reliabilities are seldom reported in studies using multiple measures of rating accuracy, it is unknown whether or not this is a characteristic phenomenon.

Rating error. A significant effect for training on halo was found, $F(2, 99) = 5.63$, $p < .01$, $sR^2 = .10$. Newman-Keuls analyses indicated that, as expected, the FOR group's ratings demonstrated significantly less halo than did either the INFO or NOT group's. The fact that the mean halo scores for all three groups were positively signed (see Table 1) indicates that, in general, subjects' ratings were characterized by slightly less variability than target scores. No significant effect of training was found on leniency-severity.

Arousal. A significant effect of training on perceived feelings of anxiety was found, $F(2, 99) = p > .05$, $sR^2 = .06$. Newman-Keuls analyses revealed that the FOR group reported significantly greater perceived feelings of anxiety than did either the INFO or NOT group. No difference existed between the INFO and NOT groups on this variable. Also, no significant effect of training was found on thought interference. This result indicates that FOR training is more arousing than either INFO or NOT training.

Group Size Effects

The means and standard deviations for all dependent variables by group size are presented in Table 2. Contrary to the hypothesized effect, group size had no significant main effect either on the retention of training information or on any of the

Table 2
Means and Standard Deviations of Dependent Variables by Group Size

Variable	<i>n</i> = 1		<i>n</i> = 6		<i>n</i> = 12	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Feelings of anxiety	17.11	4.64	15.14	5.74	15.28	5.35
Thought interference	18.83	5.87	18.47	6.21	18.50	6.07
Training retention	6.06	3.47	6.08	2.50	5.92	2.81
Orientation retention	3.47	1.25	3.31	0.98	2.47	1.25
Distance accuracy	15.95	4.18	16.05	3.77	17.27	5.06
Correlation accuracy	.50	.20	.54	.16	.51	.23
Leniency/severity	8.12	7.38	8.38	7.49	8.77	8.35
Halo	1.91	0.85	1.91	0.79	1.87	0.76

measures of rating accuracy and error within the FOR and INFO groups. Also, no significant main effect of group size was found on either arousal measure.

There was a significant group size effect, however, on the retention of pretraining information $F(2, 99) = 7.36, p < .001, \eta^2 = .15$. Newman-Keuls analyses revealed that the 1-member and 6-member groups both remembered significantly more pretraining orientation information than did the 12-member groups. No significant difference existed between the 1-member and 6-member groups on this variable. These analyses suggest that group size does not affect the retention of rater training information and the accuracy of subsequent ratings. However, subjects in the large groups ($n = 12$) do not retain as much pretraining information regarding the nature and purpose of the experiment as do those in much smaller groups ($n = 1, n = 6$). These results suggest that a social facilitation effect may have occurred during the initial stages of the experiment only.

No significant interactions of training with group size were observed on any of the dependent variables.

Discussion

The results of this study support the hypothesis that FOR training facilitates better learning of training information than does INFO training. Because the only difference between these two training approaches lies in the provision of rating standards and behavioral examples of rating dimensions in the FOR group, it is these two factors that must be responsible for the improved learning. We suggest that these findings are consistent with levels-of-processing theory (Craik & Lockhart, 1971; Jacoby & Craik, 1979). The FOR training's use of rating standards and behavioral examples represents a cognitively more elaborate and meaningful source of information for raters than does the mere presentation of rating scale items and corresponding behavioral components.

In addition, FOR training was found to result in improved distance accuracy and less halo than INFO training or no training. This result coincides with previous research that supports the superiority of FOR training over the other methods (McIntyre et al., 1984; McIntyre & Athey, 1985; Pulakos, 1984, 1986). The fact that FOR training resulted in both improved learning of training information and improved distance accuracy and halo suggests that rating accuracy is a function of the amount of rater training information that is actually retained

by raters. In support of this finding, Bernardin et al. (1985) found a weak but significant relation between accuracy of ratings and FOR knowledge. However, two findings in this study cast doubt on this conclusion. First, INFO-trained subjects retained significantly more training information than did those in the NOT group but provided ratings that were no more accurate. Second, a comparison of high-retention subjects with low-retention subjects in the INFO and FOR groups revealed no significant relation between retention and accuracy and or rating error. Other research has identified similar findings. McIntyre and Athey (1985) found that FOR subjects knew more of the training material and were more accurate in their ratings. However, no significant correlation between retention and accuracy was found. As in the present study, the lack of statistical significance of the correlation ($r = .22$) found by McIntyre and Athey may have been due to low statistical power.

A related concern is the relatively low reliability of the target scores themselves and the effect this may have had on the correlation accuracy measure. Given an obtained mean correlation accuracy value of .51 for all of the groups in this study, and interrater reliability values for the target scores and obtained ratings of .64 and .94, respectively, the corrected correlation accuracy value under conditions of perfect reliability would be .66 (see Hunter, Schmidt, & Jackson, 1982). This attenuation, in turn, diminished the magnitude of the training and group size effect on this variable. Furthermore, the observed effect size for correlation accuracy was a only .01, resulting in power of only .20 ($n = 108, k = 2$). If the effect size for correlation accuracy was raised to even .10 by improving the reliability of the target scores, the resulting power would be .80. The importance of improving the reliability of target scores in rater training research cannot be overstated.

The important point here is not to find excuses for the lack of significance of correlations and F tests. Rather, it is to recognize that a simple causal relation between rating knowledge delivered in FOR training and rating accuracy may not exist. In this regard, both this study and the one carried out by McIntyre and Athey (1985) suggest that emotional arousal may intervene to effect greater degrees of accuracy in FOR training. The FOR training provides a more interesting and challenging training experience for subjects than does INFO training and may enhance the motivation of subjects to learn the training material and perform well on the rating exercises.

Additional insight into the role that rater training plays in improving rating accuracy is gained from an item analysis of the training information retention scale. This analysis revealed that the mean retention levels for the FOR, INFO, and NOT groups were .59, .53, and .39, respectively. Because a significant effect of training on distance accuracy and halo was found in the FOR group only, it seems that substantial improvements in training retention must be achieved (e.g., an increment of at least .20 over that possessed by untrained raters) before the quality of ratings may improve significantly. Steps taken to enhance the learning of rater training content (e.g., repetition and practice) that yield only moderate improvements in actual retention level may have little effect on the accuracy of subsequent ratings. These findings also illustrate that even under the best training conditions, raters may actually retain no more than 60% of the information presented. Enhancing the retention of rater training information may, therefore, be an important ingredient in improving any rater training format.

In summary, the results of this study have a number of implications for future research on the effect of rater training on rating accuracy. First, the provision of rating standards and behavioral examples in FOR training seems to be responsible for improved rating accuracy. However, further study will be necessary before the unique effect of these two components can be fully understood. Second, the results of this study suggest that rating accuracy may not be a simple function of rating knowledge. The clarification of behavioral dimensions must be paired with semantically relevant information (e.g., rating standards, behavioral examples) to be useful to raters in evaluating performance. Further research is necessary to clarify the training information actually used by raters when they are evaluating performance and, also, to determine how rater training formats can be improved to optimize the use of this information. Finally, other links in the chain of events from rater training to rater accuracy must be investigated. In particular, it is suggested that rater arousal and motivation may play a significant role in improving the accuracy of ratings. However, the nature of this relation is unclear, indicating a need for further research on this issue. The pursuit of these questions may lead researchers to a new understanding of the role of rater training in improving rating accuracy.

References

- Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In R. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 2, pp. 230-241). New York: Academic Press.
- Beatty, M. J. (1980). Social facilitation and listening comprehension. *Perceptual and Motor Skills*, 51, 1222.
- Bernardin, H. J., Abbott, J., & Cooper, D. (1985, August). *The effects of appraisal purpose and rater training on rating characteristics*. Paper presented at the meeting of the Academy of Management, San Diego, CA.
- Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review*, 6, 205-212.
- Bloom, B. S., Engelhart, M. D., Walker, H. H., Furst, E. J., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives; Handbook 1: Cognitive domain*. New York: David McKay.
- Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology*, 60, 556-560.
- Cooper, W. H. (1981). Conceptual similarity as a source of illusory halo in performance ratings. *Journal of Applied Psychology*, 66, 302-307.
- Costin, F. (1974). Measuring lecturing behavior of college instructors. *Professional Psychology*, 1, 106-108.
- Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11, 671-684.
- Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." *Psychological Bulletin*, 52, 177-193.
- Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology*, 66, 127-148.
- Hunter, J., Schmidt, F., & Jackson, G. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.
- Jacoby, L., & Craik, F. (1979). Effects of elaboration of processing at encoding and retrieval: Trace distinctiveness and recovery of initial context. In L. S. Cermak & F. I. M. Craik (Eds.), *Human memory: A cognitive view* (pp. 1-19). Hillsdale, NJ: Erlbaum.
- Klausmeier, H. J., Wiersma, W., & Harris, C. W. (1963). Efficiency of initial learning and transfer by individuals, pairs, and quadrads. *Journal of Educational Psychology*, 54, 160-164.
- Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, 87, 72-107.
- McIntyre, R., & Athey, T. R. (1985, August). *Rater training research: Group size and other methodological considerations*. Paper presented at the meeting of the Academy of Management, San Diego, CA.
- McIntyre, R., Smith, D., & Hassett, C. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology*, 69, 147-156.
- Moeser, S. (1983). Levels of processing: Qualitative differences or task-demand differences? *Memory and Cognition*, 11, 316-323.
- Murphy, K., Garcia, M., Kerkar, S., Martin, C., & Balzer, W. (1982). Relationship between observational accuracy and accuracy in evaluating performance. *Journal of Applied Psychology*, 67, 320-325.
- Nathan, B. R., & Lord, R. G. (1983). Cognitive categorization and dimensional schemata: A process approach to the study of halo in performance ratings. *Journal of Applied Psychology*, 68, 102-114.
- Oetting, E. R., & Deffenbacher, J. C. (1980). *Test anxiety profile manual*. Ft. Collins, CO: RMBSI.
- Pulakos, E. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology*, 69, 581-588.
- Pulakos, E. D. (1986). The development of training programs to increase accuracy with different training tools. *Organizational Behavior and Human Decision Processes*, 38(1), 76-91.
- Reed, S. K. (1982). *Cognition: Theory and applications*. New York: Brooks/Cole.
- Reynolds, A. G., & Flagg, D. W. (1983). *Cognitive Psychology*. Boston, MA: Little, Brown.
- Snedecor, G. (1946). *Statistical methods*. Ames: Iowa State College Press.
- Zajonc, R. B. (1965). Social facilitation. *Science*, 149, 269-274.
- Zajonc, R. B. (1980). Compresence. In P. B. Paulus (Ed.), *Psychology of group influence* (pp. 35-60). Hillsdale, NJ: Erlbaum.

Received February 15, 1985

Revision received April 10, 1987

Accepted April 16, 1987 ■

Behavioral Anchors as a Source of Bias in Rating

Kevin R. Murphy and Joseph I. Constance
Colorado State University

Behavioral anchors may affect the way that raters process information about ratees, and may in some cases be a source of bias in rating. This study tested the hypothesis that the presence of behavioral anchors that closely matched behaviors actually observed by raters would bias performance ratings. Subjects ($N = 180$) viewed videotaped lectures and rated them, using scales that contained examples of either good or bad performance that had actually occurred on the tapes, but that were not representative of the ratee's overall performance. One half of the subjects read the scales before viewing the lectures; the remaining subjects read the scales only after viewing the lectures. There was a significant scale effect, but no Scale \times Order interaction; ratings were biased in the direction of unrepresentative anchors. These results suggest that behavioral anchors can be a source of bias in ratings and they may lead to biased recall, but they probably do not bias the observation and encoding of ratee behavior. Our results suggest that behaviorally anchored scales are not necessarily more objective or less prone to bias than are scales without behavioral anchors.

Several methods have been proposed to reduce bias in performance appraisal. One strategy is to use rating-scale formats that include concrete behavioral anchors that illustrate the performance dimensions and performance levels to be rated (e.g., behaviorally anchored rating scales [BARS], Smith & Kendall, 1963; behavior observation scales, Latham & Wexley, 1977; and mixed standard rating scales, Blanz & Ghiselli, 1972). Although rating scale formats are thought to have little effect on the psychometric properties of ratings (Landy & Farr, 1980), the use of behavior-based rating scales may be worthwhile because concrete behavioral anchors should reduce unnecessary subjectivity in rating (see Murphy, Martin, & Garcia, 1982). However, when behavior-based scales are examined from an information-processing perspective, it is not clear whether the inclusion of specific behavioral anchors will contribute to or hinder the accuracy of performance evaluations.

Several authors have suggested that scale formats may affect the cognitive processes involved in performance rating (Atkin & Conlon, 1978; Bernardin & Smith, 1981; DeNisi, Cafferty, & Meglino, 1984; Schwab, Heneman, & DeCotiis, 1975). Although behavioral anchors may affect the encoding and retrieval of behavioral information, behavior-based scales do not necessarily contribute to accuracy in rating; behavioral anchors may interfere with rather than enhance the accurate processing of behavioral information. This study tests the hypothesis that behavioral anchors may serve as a source of rating bias by distorting the observation or recall of ratee behavior.

Anchors as a Source of Bias

Behavioral anchors provide the rater with a standard framework for evaluating performance and help to illustrate the types

of behaviors that might be expected from good, average, and poor performers. In some cases, the behaviors included on rating forms are highly similar to specific behaviors observed by the rater; the presence of such anchors could affect the cognitive processes of the rater. First, the rater's attention may be directed toward behaviors included on rating scales, thus increasing the salience of those behaviors and decreasing the relative salience of behaviors not included on the rating form. Second, these behavioral anchors may guide the rater's memory for performance-relevant information. Behaviors listed on rating forms may be more easily remembered than behaviors not listed on the forms.

Behavioral anchors may bias ratings if they misdirect the rater's observation or recall of ratee behavior. In particular, behavioral anchors present a potential problem if they describe behaviors that are actually observed by raters but that are not representative of the ratee's overall performance levels. For example, a truly good performer will sometimes exhibit ineffective behaviors. If behavioral anchors direct the rater's attention to or facilitate the recall of those unrepresentative behaviors, ratings of that person may be unfairly low. Thus, the use of behavioral anchors may in some cases lead to rating bias.

Biases in Observation

Behavioral anchors could affect the rater's observation of ratee behavior in several ways. First, the inclusion of specific behaviors on a rating form could increase the likelihood that raters will recognize and attend to those behaviors. Indeed, one advantage of behavioral anchors is that they help to identify behaviors regarded by the employer as important (Landy & Farr, 1983). Another possibility is that anchors may indirectly affect behavior observation by priming those categories of behaviors that are related to the anchors included on a rating scale (Higgins, Rholes, & Jones, 1977). A rater who has examined a behaviorally anchored rating scale may pay more attention to behaviors similar to those present on the scale than to those that

This article is based on research presented at the conference on Decision Making and Information Processing, held at the State University of New York at Buffalo, October 1986.

Correspondence concerning this article should be addressed to Kevin R. Murphy, Department of Psychology, Colorado State University, Fort Collins, Colorado 80523.

are not, and may interpret ambiguous behaviors in terms similar to the anchors he or she has read. For example, if a rating scale includes as an anchor "Employee insults customers," a rater who is familiar with this rating scale may perceive any lack of deference to customers on the part of his or her subordinate as an insult and as a sign of very poor performance.

Observational biases may distort ratings even when all of the behaviors included on the form are valid indicators of good or poor performance. Consider, for example, a worker who generally is neither a good nor a poor performer, but who once in a great while exhibits behaviors normally expected of very superior workers. If the rater's attention has been directed by behavioral anchors toward those specific behaviors, the worker's overall performance may be perceived as much better than it actually is.

Bias in Recall

In most performance appraisals, raters must remember a great deal of information about each ratee. The presence of behavioral anchors may change performance rating from a free-recall task to a cued-recall task. In particular, if *actually observed* behaviors closely match behavioral examples included as anchors on the rating form, the presence of such anchors will facilitate the memory of those behavioral incidents (Tulving, 1983). Research on the availability heuristic (Kahneman & Tversky, 1973, 1982; Tversky & Kahneman, 1974) suggests that the ability to easily remember specific examples of good or bad performance will bias ratings in the direction of those examples.

As we stated in the preceding section, behavioral anchors can distort the retrieval of performance-related behavior if incidents similar to those included on the form, but not representative of the ratee's overall performance level, have been observed by the rater.

Hypotheses

Bias

Subjects viewed and rated videotaped lectures, using behaviorally anchored rating scales constructed for this study. The scales included as anchors specific behavioral incidents that had occurred on the tapes but that were not representative of the ratee's overall level of performance. Specifically, the videotapes used in this study featured lectures that were generally of average quality (neither clearly good nor bad), but that also featured several specific behaviors that exemplified both very good and very poor performance. Our first hypothesis was the following: The use of scales that include as anchors unrepresentative behaviors that are observed by the rater will bias rating in the direction of those anchors.

Observation Versus Recall

To determine whether behavioral anchors would bias either behavior observation or memory for behavior, or both, we varied the order of videotape and the rating scale presentations. One half of the subjects examined the rating scales prior to viewing the tapes (previewing condition); the remaining sub-

jects were not shown the scales until after they had viewed the tapes (postviewing condition). In the previewing condition, anchors can affect the observation of ratee behavior, memory for ratee behavior, or both. However, in the postviewing condition, anchors cannot affect behavior observation because anchors are not seen until after the ratees are observed, but anchors may affect rater's retrieval of behavioral information. Thus, a comparison between ratings obtained in these two conditions could help to isolate the effects of anchors on observation and memory. For example, if anchors bias both observation and memory, a larger anchor effect will be found in the previewing condition than in the postviewing condition, in which observation biases are ruled out. If anchors bias observation, but not memory, an anchor effect will be found in the previewing condition but not in the postviewing condition. Finally, if anchors bias memory but not observation, anchor effects of equal magnitude will be found in the pre- and postviewing conditions.

Our second hypothesis was that there would be an interaction between the order of viewing the videotapes and scales, and the type of anchors included on the rating scale: The effect of behavioral anchors will be different when scales are examined prior to observing ratee behavior than when they are examined only after observing ratee behavior. No specific form was hypothesized for this interaction.

Method

Subjects viewed videotaped lectures that depicted an average level of performance but that included a few specific behaviors associated with both very good performance and very poor performance, and rated them using behaviorally anchored rating scales. We developed rating scales that included as anchors either specific examples of good performance that had occurred on the tapes (high anchor condition), specific examples of poor performance that had occurred on the tapes (low anchor condition), or rating scales that did not include as anchors any of the specific behaviors that actually occurred on the tapes (control group). Subjects were shown these scales either before or after viewing the videotapes.

Videotapes

Videotapes developed by Murphy, Balzer, Lockhart, and Eisenman (1985) were used in this study. In the specific set of tapes used here, an actress played the role of a graduate student lecturing on the topic of sleep. Lectures were carefully scripted to portray an average (i.e., neither unusually good nor bad) level of performance, but also included specific behaviors associated with both very good and very poor performance.

To verify that the tapes used here were perceived by subjects as average, 21 undergraduates were asked to rate the tapes on three dimensions: quality of the lecture, response to questions, and speaking style. They used a 9-point scale, on which a rating of 1 indicated *very poor performance*, 9 indicated *very good performance*, and 5 indicated *average performance*. The mean ratings on these three dimensions were 5.4, 5.5, and 4.7, respectively, indicating that when behaviorally anchored scales are not used, the tapes are regarded as average in quality.

Scale Development

The authors reviewed the videotapes and recorded 32 behavioral incidents that reflected good, average, or poor performance. Another set of 19 incidents that were plausible but that had *not* occurred on the tapes

was also generated. This set of 51 items was presented in random order to separate groups of undergraduate ($n = 24$) and graduate ($n = 12$) psychology students. Subjects were asked to classify each incident into one of the three aforementioned rating dimensions and to rate the level of performance exemplified by each incident, using a 9-point scale. This procedure, similar to that used in developing BARS (Landy & Farr, 1983), allowed us to identify behavioral anchors that had either occurred or not occurred on the tapes and that were clear exemplars of good, average, and poor performance on each of the three rating dimensions.

Behaviorally anchored rating scales were developed for the three rating dimensions (quality of the lecture, response questions, and speaking style). Each scale included one behavioral anchor illustrating good performance, one illustrating poor performance, and one illustrating average performance. Three separate BARS were developed for each dimension. In the control group, none of the anchors included on the form were behaviors that had actually occurred on the tapes. In the high anchor condition, the behavioral anchor illustrating a high level of performance on each dimension described an incident that had actually occurred on the tapes but that was unrepresentative of the lecturer's overall performance. The anchors used to describe average and poor performance were the same as those used in the control group. In the low anchor condition, the behavioral anchors illustrating a low level of performance on each dimension described incidents that had actually occurred on the tapes. The anchors used to describe good and average performance were the same as those used in the control group. The scales used for each dimension in each of the three conditions are shown in Figure 1.

The anchors enclosed in boxes in Figure 1 represent behaviors that actually occurred on the tapes; behaviors not enclosed in boxes did not occur on the tapes. Thus, our manipulation of the behavioral anchors consisted of either presenting scales that did not list any of the behaviors shown in boxes in Figure 1 (control group), scales that substituted the boxed example (behaviors that actually occurred) of good performance for the unboxed example (high anchor condition), or scales that substituted the boxed example of poor performance for the unboxed example (low anchor condition).

Subjects

A total of 180 undergraduates participated in this study in exchange for course credit. This number of subjects provides statistical power in excess of .80 for detecting main effects and in excess of .70 for detecting interactions, assuming a small effect in the population.

Procedure

Groups of from 6 to 10 subjects were randomly assigned to one cell of a 3×2 factorial experiment, in which the variables manipulated were scale (high anchor, low anchor, control) and order (previewing, postviewing). Subjects were told that their task was to watch videotaped lectures and to rate the lecturer's performance. Subjects in the previewing conditions were then shown the scales they would be using and were encouraged to read the behavioral anchors. After examining the scales, subjects viewed the videotapes and immediately rated the lecturer using the BARS provided by the experimenter. Subjects in the postviewing condition viewed the tapes first and were shown the rating scales immediately after they had seen the tapes. After reading the scales, subjects rated the lecturer.

Subjects in the high anchor condition received BARS for each of the three rating dimensions that included, as their illustration of good performance, a behavioral incident that had actually occurred on the tape. Similarly, subjects in the low anchor condition received scales for each dimension that included as their illustration of poor performance be-

havioral incidents that had actually occurred on the tape. None of the behavioral anchors on the scales used by the control group described incidents that had actually been observed by raters.

Results

We used a factorial multivariate analysis of variance to test the hypothesis that variation in behavioral anchors would bias ratings. We found a significant scale effect, $F(6, 344) = 7.57$, $p < .01$, but no Scale \times Order interaction. We did not hypothesize a main effect for order, nor was one found. Follow-up analyses of variance, presented in Table 1, suggest that variation in the anchors used in measuring quality of the lecture, and speaking style had a significant impact on ratings. The means shown in Table 2 indicate that the effects were in the predicted direction. That is, ratings in the high anchor condition were higher than in the control group; ratings in the low anchor condition were lower than in the control group.

We used Dunn's procedure to test these contrasts and found significant ($p < .05$) differences between all three means for the quality of lecture scale. The high anchor mean was significantly larger than the low anchor mean for speaking style, but other contrasts were not significant.

Discussion

Our results confirm the hypothesis that behavioral anchors can be a significant source of bias in performance ratings. In particular, when BARS contained incidents that had actually been observed by the rater, but that were not representative of the ratee's performance, performance ratings were biased in the direction of those unrepresentative anchors. In this study, behavioral anchors, which are usually recommended as a method of decreasing subjectivity and bias, were themselves a source of bias in performance rating.

Earlier, we noted that behavioral anchors might bias observation or memory, or both. Our results suggest that anchors do not bias observation but may bias memory for behavior. First, we found a main effect for scale, but no Scale \times Order interaction. In other words, anchors had the same effect in the postviewing condition as in the previewing condition, suggesting that biases in observation are not necessary to account for the effects of behavioral anchors. To provide a more definitive test of the hypothesis that observation biases are not the source of the behavioral anchor effect observed here, we analyzed separately the data from our postviewing condition, in which observational biases can be clearly ruled out. We found a significant multivariate scale effect, $F(6, 172) = 4.39$, $p < .01$, in the postviewing condition; as in Table 2, the mean ratings on all three scales were in the predicted direction.

Our results suggest that behavioral anchors bias the retrieval of behavioral information but are not definitive in this regard. For example, it is possible that differences in scale anchors affect the raters' perception of the scale used (Ironson & Smith, 1981). That is, a rating of 6 on a 9-point scale could be interpreted as either fairly average or definitely above average, depending on the anchor used to illustrate a 6. Thus, it is possible that the different anchors used in the high anchor, low anchor, and control group lead to slightly different interpretations of scale val-

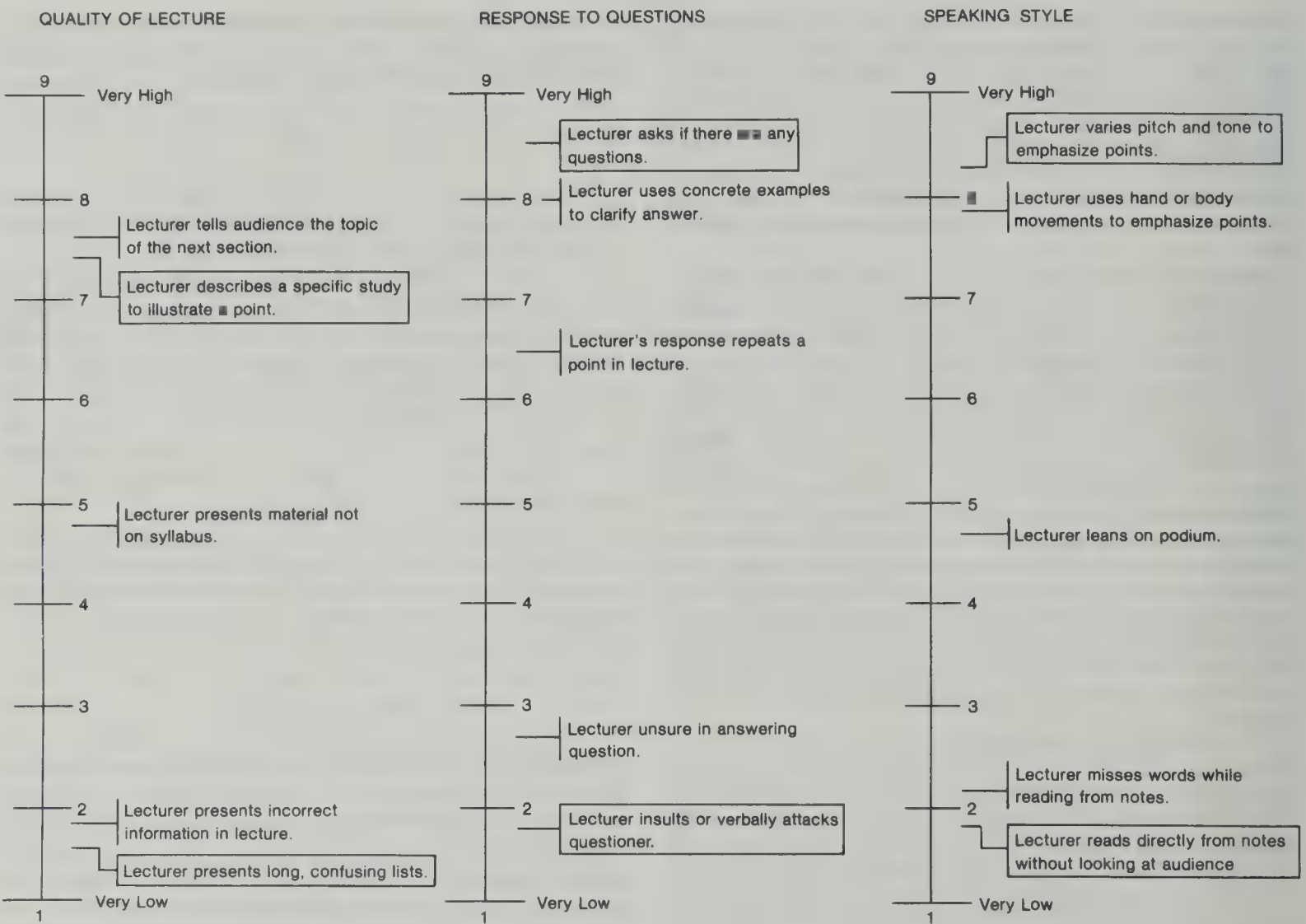


Figure 1. Rating scales for high anchor, control, and low anchor conditions. (High and low anchors are enclosed in boxes.)

ues. It is also possible that variation in scale anchors changes strategies for integrating information rather than changing behavioral memory. That is, subjects in our high and low anchor conditions may have remembered the same behaviors as those in the control group, but may have given more weight to behav-

Table 1
Effects of Scale Differences and Order for Three Performance Dimensions

Effect	df	Quality of lecture		Response to questions		Speaking style	
		F	ω^2	F	ω^2	F	ω^2
Scale ^a	2	21.19*	.23	1.91	.01	6.82*	.08
Order	1	0.19	.00	2.39	.02	3.52	.02
Scale \times Order	1	1.70	.01	0.19	.00	0.71	.00
Error	175						

^a High anchor; control; low anchor.
* $p < .01$.

iors that were specifically mentioned as scale anchors than to those that were not. A more sensitive test of the memory hypothesis would require variation in the delay between observing ratee behavior and examining the behaviorally anchored scales used in this study.

The usual strategy for developing BARS is to develop clear, valid examples of good, average, and poor performance. Anchors that reflect critical behaviors, either good or bad, are thought to be especially useful (Latham & Wexey, 1977). The results of this study suggest that the representativeness of these behaviors must also be considered. If behavioral anchors do in

Table 2
Means and Standard Deviations for Three Rating Scales

Rating dimension	High anchor		Control		Low anchor	
	M	SD	M	SD	M	SD
Quality of lecture	6.08	1.41	5.10	1.44	4.20	1.84
Response to questions	5.34	1.68	5.10	1.71	4.72	1.86
Speaking style	4.36	1.66	3.62	1.33	3.43	1.38

some way guide the rater's memory for performance-related information, anchors should be selected that have the greatest likelihood of facilitating memory of representative rather than atypical behaviors. For example, if you were developing a BARS for shipping clerks, an example such as "Routes shipment to the wrong country" might be a clear example of spectacularly poor performance, but is probably not representative of the typical poor performer. It is likely that poor performers make large numbers of minor errors rather than one spectacular error; a clerk who was generally quite efficient, but who had once routed a shipment to the wrong country, would probably receive a lower rating on this BARS than would a truly poor performer.

This study adds to the growing body of evidence that suggests that behaviorally oriented scales are not always more objective than trait scales (Kavanaugh, 1971; Murphy et al., 1985; Murphy et al., 1982). To date, behavior-based appraisal systems have been accepted at face value. For example, a performance appraisal system that involves specific behavioral information is generally accepted as more credible in equal employment litigation than a simple graphic scale (Cascio & Bernardin, 1981). If behavioral anchors distort rather than enhance raters' ability to remember performance-related information, it is possible that behavior-based scales will provide *less* accurate measurement of performance than will simple graphic scales.

Smith and Kendall (1963) developed behaviorally anchored scales as a method of providing a standard framework for observing and recording behaviors. Although the procedures outlined by Smith and Kendall (1963) are rarely followed when using BARS-like scales (Bernardin & Smith, 1981), our results are consistent with many of their original recommendations. First, anchors did not bias behavior observation, only memory for behavior. Thus, if BARS are used as aids for observing and recording behavior, the sort of bias observed in this study is unlikely to occur. Second, Smith and Kendall (1963) recommended that anchors represent typical behaviors, or those that could be expected. Although we considered the possible effects of behavioral anchors from a very different perspective, our recommendation that anchors be representative is not very different from theirs. Thus, the results of our study do not suggest that the BARS technique be abandoned. Rather, they suggest that greater care must be exercised in constructing, using, and evaluating BARS.

References

- Atkin, P. S., & Conlon, E. J. (1978). Behaviorally anchored rating scales: Some theoretical issues. *Academy of Management Review*, 3, 119-128.

- Bernardin, H. J., & Smith, P. C. (1981). A clarification of some issues regarding the development and use of behaviorally anchored rating scales (BARS). *Journal of Applied Psychology*, 66, 458-463.
- Blanz, F., & Ghiselli, E. E. (1972). The mixed standard scale: A new rating system. *Personnel Psychology*, 25, 185-200.
- Cascio, W. F., & Bernardin, H. J. (1981). Implications of performance appraisal litigation for personnel decisions. *Personnel Psychology*, 34, 211-226.
- DeNisi, A. S., Cafferty, T. P., & Meglino B. M. (1984). A cognitive view of the performance appraisal process: A model and some research propositions. *Organizational Behavior and Human Performance*, 33, 360-396.
- Higgins, E. T., Rholes, W., & Jones, C. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology*, 13, 114-154.
- Ironson, G. H., & Smith, P. C. (1981). Anchors away: The stability of meaning of anchors as their location is changed. *Personnel Psychology*, 34, 249-262.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237-251.
- Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition*, 11, 123-141.
- Kavanaugh, M. (1971). The content issue in performance appraisal: A review. *Personnel Psychology*, 34, 653-668.
- Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, 87, 72-107.
- Landy, F. J., & Farr, J. L. (1983). *The measurement of work performance: Methods, theory, and applications*. New York: Academic Press.
- Latham, G. P., & Wexley, K. N. (1977). Behavioral observation scales for performance appraisal purposes. *Personnel Psychology*, 30, 225-268.
- Murphy, K. R., Balzer, W. K., Lockhart, M. C., & Eisenman, E. J. (1985). Effects of previous performance on evaluations of present performance. *Journal of Applied Psychology*, 70, 72-84.
- Murphy, K. R., Martin, C., & Garcia, M. (1982). Do behavioral observation scales measure observation? *Journal of Applied Psychology*, 67, 562-567.
- Schwab, D., Heneman, H. G., & De Cotiis, T. (1975). Behaviorally anchored rating scales: A review of the literature. *Personnel Psychology*, 28, 549-562.
- Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 47, 149-155.
- Tulving, E. (1983). *Essentials of episodic memory*. New York: Oxford University Press.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.

Received December 12, 1986

Revision received February 10, 1987

Accepted April 3, 1987 ■

Measuring Occupational Difficulty: A Construct Validation Against Training Criteria

Michael D. Mumford
Georgia Institute of Technology

Joseph L. Weeks
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

Francis D. Harding
Advanced Research Resources Organization
Bethesda, Maryland

Edwin A. Fleishman
George Mason University

Burtch, Lipscomb, and Wissman's (1982) occupational learning difficulty index attempts to measure the difficulty of occupations by aggregating workers' evaluations of task learning time. In the present study we examined the construct validity of this job analysis index. To accomplish this, 48 different occupational training programs were described in terms of 15 training content variables, 6 student characteristics variables, and 7 training performance variables. The results, obtained in a correlational analysis, indicated that the occupational learning difficulty index yielded an interpretable pattern of relationships with the training content and performance variables. We conclude that this task learning time index displays some construct validity as a measure of occupational difficulty and, therefore, should prove of value in designing training, manpower allocation, and job evaluation systems.

Personnel administrators are often faced with decisions that require an assessment of occupational difficulty. For instance, in manpower allocation systems, attempts are frequently made to assign the most talented individuals to the most difficult occupational specialties (Weeks, 1984). Further, occupational difficulty is often an important issue in the design of training programs as well as in decisions concerning pay rates and comparable worth (Cronbach & Snow, 1977; McCormick, 1979). Given the fundamental importance of these decisions, there would seem to be a need for an objective measure of occupational difficulty.

Because manpower allocation and training design problems are especially salient in the military context, the Air Force has devoted substantial effort to the measurement of occupational difficulty. In part, this effort was spurred by the logical inconsistencies resulting from the use of historic attrition rates and manpower requirements, rather than an objective measure of occupational difficulty, for solving these problems (Weeks, 1984). Initial Air Force research into the measurement of occupational difficulty was based on incumbent judgments of task difficulty. However, this approach was not successful due to the confounding of task difficulty characteristics with immediate climatic influences such as time pressure and leader-subordinate relationships (Madden, 1962).

A potential solution to this problem was developed on the

basis of certain findings in the educational research literature. Essentially, it was argued that a more difficult job task or activity would be one that took individuals a longer time to master with ability held constant (Cronbach & Snow, 1977; Gettinger & White, 1979; Krumboltz, 1965). Research that applied this principle within an occupational field demonstrated that incumbents in different work situations displayed substantial agreement in their evaluations of the time required to learn various tasks (Leczner, 1971; Mead & Christal, 1970). Furthermore, these evaluations of task learning time were found to be related to perceived task aptitude requirements (Fugill, 1972). On the basis of evidence provided by these studies, Christal (1974) concluded that a measure of occupational difficulty might be formulated, first, by having an adequate sample of judges evaluate the time required to learn each task performed in an occupational specialty and, then, by summing these average ratings across all tasks performed in a particular occupation, after weighting for the time devoted to task performance.

Burtch, Lipscomb, and Wissman (1982) attempted to implement this task learning time approach to the measurement of occupational difficulty by using a series of benchmark rating scales. These rating scales were developed by having job analysts and incumbents evaluate the relative learning time of a sizable sample of well-known and easily observable tasks. Two tasks were then selected to represent each interval on a 25-point scale. The learning time of new tasks was established by reference to these task scale points. Occupational learning difficulty was defined by weighting these learning time ratings for total time spent in task performance, and then aggregating these weighted values across all tasks performed in an occupation.

Although subsequent research has demonstrated the reliability and administrative feasibility of this technique (Burtch et al., 1982), evidence pertaining to the validity of such occupa-

The authors would like to note that this work was carried out under research Contract F33615-83-C-0036 awarded to the Advanced Research Resources Organization by the Air Force Human Resources Laboratory.

Correspondence concerning this article should be addressed to Michael D. Mumford, School of Psychology, Georgia Institute of Technology, Atlanta, Georgia 30332.

Table 1
Occupational Specialties Sampled

Title	Title
Aircraft Environmental Systems Mechanic	Continuous Photo Processing Specialist
Aircraft Fuel Systems Mechanic	Instrumentation Mechanic
Airframe Repair Specialist	Avionic Sensor Systems Mechanic
Missile Maintenance Specialist	Precision Measuring Equipment Specialist
Special Vehicle Mechanic	Fire Protection Specialist
Aircraft Electrical Systems Specialist	Aerospace Ground Equipment Specialist
Jet Engine Mechanic	Computerized Test Station Specialist
Aircraft Loadmaster	Attack Control Systems Specialist
Telephone Switching Specialist	Munitions Systems Specialist
Cable Splicing Specialist	Material Facilities Specialist
Tactical Aircraft Maintenance Specialist	Armament Systems Specialist
Electrician	Command and Control Specialist
Carpentry Specialist	Wideband Equipment Specialist
Financial Services Specialist	Electronic Computer Specialist
Medical Services Specialist	Telecommunications Control Specialist
Surgical Services Specialist	Airborne Warning Radar Specialist
Medical Assistant Specialist	Electronic Warfare Systems Specialist
Physical Therapy Specialist	Computer Specialist
Dental Assistant Specialist	Administration Specialist
Financial Management Assistant	Personnel Specialist
Medical Laboratory Specialist	Personal Affairs Specialist
Security Specialist	Ground Radio Operator
Law Enforcement Specialist	Aircraft Warning Radar Specialist
Navigation Systems Specialist	Computer Programming Specialist

tional difficulty measures has not been obtained. However, the potential impact of this information on personnel policy underscored the need to gather some validation evidence prior to routine implementation. Thus, the present study represents an attempt to establish the construct validity of this occupational learning difficulty measure in a sizable sample of Air Force specialty fields.

Method

Sample

The sample of jobs used in the present investigation consisted of 48 distinct entry-level occupational specialties drawn from a total population of some 200 entry-level specialties that exist in the Air Force (Weeks, 1984). These occupations were chosen for study by Air Force administrators and personnel researchers to ensure diversity in job content as well as representativeness with respect to training content, training costs, training center location, yearly flow through a training course, and minimum aptitude requirements for entry into occupational training programs. Table 1 presents titles of the 48 occupational specialties chosen for study.

Because certain measures used in this investigation reflected aggregates of individual data, we had to obtain an adequate sample of individuals assigned to each occupational specialty. Consequently, within each occupation, all of the individuals who entered the relevant specialty training course between June and December of 1983 were sampled. This procedure provided a total of 5,970 subjects, with an average of 127 subjects and a minimum of 20 subjects per occupational specialty. Most sample members were White men; their average age was 20 years; and almost all of them were high school graduates whose scores on the Armed Forces Qualifying Test exceeded 50. Note that these demographic characteristics are typical of the Air Force enlistee population, and so point to the adequacy of the sampling procedures we used.

Occupational Difficulty

The relative difficulty of the 48 occupational specialties included in this sample was established by using evaluations of task learning time from the Burtch et al. (1982) benchmark scales. These scales describe task learning time for occupational specialties within the general, administrative, electronics, and mechanical aptitude areas, and were constructed in accordance with the following format. Initially, 600 well-known and easily observed tasks having different performance requirements were drawn from the occupational specialties in each of the four aptitude areas. In each aptitude area, 8 to 14 job analysts were asked to rank order these 600 tasks in terms of learning time, and 50 senior incumbents, drawn from 15 occupational specialties incorporated in the aptitude area, were asked to rate the learning time of these tasks on a 9-point rating scale. Within each aptitude area, the job analysts' distribution of learning time rankings was divided into 25 equal intervals. We then obtained the means and standard deviations of rankings of job analysts and senior incumbents. This information was used to select two tasks in each interval with means near the interval midpoint and small standard deviations. These two tasks then served as anchors for that point on the benchmark scale that applied to specialty fields in a given aptitude area. An example of the resulting rating scales may be found in Table 2, which presents the benchmark scale developed for mechanical occupations.

Application of these scales in the assessment of occupational learning difficulty begins with a task-level assessment. Initially, a list of all of the tasks currently performed in a specialty field at all levels of seniority is presented to from 10 to 20 job analysts and from 50 to 100 senior incumbents who have more than 10 years of experience. Members of both of these groups then rate the learning time of each task. Because senior incumbents are responsible for supervising the work of their less experienced juniors, they generally have an intimate familiarity with all of the tasks appearing in an occupational field within a few months of their introduction.

The foregoing procedures allow interrater agreement and conver-

Table 2

Mechanical Benchmark Scale for Learning Time Levels

Level	Task item	Level	Task item
1	Police grounds for litter	15	Perform preoperational inspections of engine after engine has been on long standby
2	Police open storage areas		
2	Cut weeds		
3	Dispose of rags		
3	Lubricate cables		Install or replace formica on counter tops or splash boards
4	Rake bar screen		
4	Lubricate hand tools	16	Remove or install canopy hoses or tubing
	Stencil date of inspection on life rafts		
5	Clean life preservers		Prime and bleed fuel systems
	Dig ditches by hand	17	Remove or replace transmission-driven generators
6	Clean paint		Adjust automatic governors and voltage regulators
	equipment such as brushes or rollers		
	Apply reflective tape to equipment	18	Troubleshoot high or lube oil pressure
7	Remove or replace venetian blinds		
	Clean equipment or areas after applying protective coatings	19	Install fuel manifolds and fuel nozzles
8	Maintain tool cribs		
	Mix concrete by hand	20	Install electrical components
9	Position nonpowered ground equipment around aircraft		Remove or install fuel cells
	Apply enamels to surfaces using rollers	21	Read and interpret schematic or wiring diagrams
10	Clean and regap spark plugs		
	Caulk areas around windows, sink, or bathtubs		Install tail rotor assemblies on helicopter aircraft
11	Perform operator inspections or maintenance on dump trucks	22	Remove or install tail drive assembly
	Drain engine oil systems		Direct aircraft explosive hazard render safe procedures
12	Remove or replace nozzles or hoses on motor gasoline units	23	Perform critical measurements on jet engines
	Prepare enamels for application		Adjust canopies
13	Install or replace water fountains		Remove or replace cyclic control system components
	Disassemble or clean conventional fuel gate valves		Remove or install main rotor transmission
14	Prime components such as starters and hydraulic pumps	24	Troubleshoot fully articulated rotor systems and determine corrective actions
	Disassemble or clean conventional fuel lubricated plug valves		Assemble main engine sections
		25	Troubleshoot systems for breaker trip-outs
			Troubleshoot installed engines

Note. Taken from Burtch, Lipscomb, and Wissman (1982).

gence among rater types to be established. Consequently, tasks associated with low interrater agreement within a type or limited convergence across rater types are submitted to further scrutiny. Typically, more than 90% of the tasks incorporated in an occupational field survive screening, and relative learning time is defined in terms the average rating of the tasks across raters and rater types.

Once the average learning time of job tasks has been determined, the occupational tasks being performed in entry-level positions are specified on the basis of existing job analysis information. The amount of time occupants of entry-level positions spend on each of these tasks is also determined, and an average of these values is obtained across raters. These average time on task figures were derived from United States Air Force Occupational Measurement Center job analysis surveys. These surveys generally yield a 75% return rate in a sample of all incumbents in occupational fields employing less than 3,000 individuals, and a stratified random sampling of more populous occupational fields.

To generate occupational learning difficulty values for each specialty, the average time on task figures are multiplied by the average learning time rating obtained for the entry-level occupational tasks. The sum of these values across all entry-level tasks performed in a specialty field provided the overall occupational learning difficulty (OLD) values used in the present study. A detailed examination of the reliability of these scales by Burtch et al. (1982) indicates that 13 job analysts rating 60 select tasks is sufficient to produce average interrater agreement coefficients in the low .90s. Furthermore, when independent evaluations made by two teams of raters were used to establish the OLD values for 93 specialty fields, a short-interval retest coefficient in the low .80s was obtained. Finally, note that general, mechanical, administrative, and electronics aptitude areas were considered equivalent for the purposes of the present study on the basis of the high correlation of scale scores. Also, the OLD values used were restricted to entry-level positions, inasmuch as those are the positions of primary concern in Air Force technical training.

Validation Strategy

Because the OLD index measures occupational difficulty in terms of task training time, it seemed desirable to validate this measure with respect to the characteristics of entry-level training programs in the 48 specialty fields. Application of this strategy appeared especially appropriate because this task time learning information was not considered in the design of training programs at the time these ratings were obtained, although Air Force training programs are carefully developed in relation to current job content and performance requirements. Moreover, it appears that these OLD ratings are not markedly influenced by occupational stereotypes. For instance, Weeks (1984) noted that many Air Force personnel do not view jet engine mechanics as a particularly demanding specialty field, although it yields one of the highest OLD values found in the Air Force. Additionally, it has been found that OLD values yield only weak positive correlations, on the order of .10, with minimum aptitude requirements for entry into specialty training programs established on the basis of administrators' a priori beliefs concerning the difficulty and importance of various occupational fields (Mumford, Harding, Fleishman, & Weeks, 1987).

Aside from these purely technical considerations, it appeared that the use of training program information permitted a far more cohesive construct validation effort than might otherwise be possible. First, description of entry-level training programs made possible identification of a set of variables relevant to description of the training process per se without reference to the OLD index. Thus, the possibility of implicit structural biases in variable selection could be avoided. Second, it facilitated the development of criterion measures applicable across a variety of occupational fields, inasmuch as Air Force entry-level training programs are carried out within a common administrative framework.

Third, because the OLD index explicitly focuses on task training time, it appeared that use of this strategy provided a heuristically appropriate framework for the construct validation effort that would facilitate hypothesis specification.

To identify variables capable of describing the training process across specialties and to generate adequate measures of these variables, a series of interviews were conducted with trainers, training supervisors, instructional design personnel, measurement and evaluation specialists, and administrative personnel at four major Air Force technical training centers. Based on the interview data, three general categories of variables were identified as being capable of influencing the training process across a number of different occupational specialties.

The first category consisted of course content variables describing general parameters of concern in instructional systems design. Variables in the second category reflected student characteristics likely to influence training performance. The third category included variables reflecting alternative measures of training performance common to all specialties. An expert content review of these three categories and their associated variables and measures, indicated that they provided an adequate description of the more significant aspects of entry-level technical training in the Air Force. Thus, it appeared that these criterion variables supplied a sufficiently comprehensive framework for drawing some general conclusions concerning the construct validity of the OLD.

Criterion Measures and Hypotheses

Table 3 presents a brief listing of the 15 course content variables Air Force personnel believed to be of some general significance in describing the characteristics of entry-level training programs. Also presented are the measures of these variables used in the present study. For the most part, the information presented in Table 3 is self-explanatory, although there are a few points worthy of further consideration.

The course length, diversity, practice, expected attrition, instructional aids, hands-on practice, feedback, yearly flow, and day-length measures were all drawn from the 1983 Training Plan (TP) and Program of Instruction (POI) of the 48 courses. Because these TPs and POIs are rigidly adhered to within the Air Force instructional system, they could be assumed to provide a highly accurate and reliable description of course content. This also seems to hold true for the instructor experience and personnel requirements measures that were obtained from Air Force administrative records applicable during the summer of 1983. On the other hand, the reading difficulty, abstract knowledge requirements, and instructional quality measures were generated on the basis of rating algorithms whose reliability could not be assumed. However, Mumford et al. (1987) have provided evidence indicating that these three measures all yield interrater reliabilities in excess of .70.

A brief description of the a priori hypotheses formulated with respect to the relations observed between the OLD index and these 15 course content variables can also be found in Table 3. As one can see, it proved possible to formulate relatively unambiguous hypotheses for 12 of the 15 course content variables. More specifically, because the OLD index is derived from the number of difficult tasks performed in an occupation in which difficulty is defined in terms of learning time, we assumed that both the course length and diversity measures would yield a strong direct relation with OLD scores. In the sense that the abstract knowledge requirements, reading difficulty, and expected attrition variables reflect the difficulty of training materials, and the difficulty of such materials should be directly related to the difficulty of job tasks, it did not seem unreasonable to hypothesize that scores on these measures would produce a significant positive relation with the OLD index.

We also held that increasing occupational difficulty would require a greater investment in effective instruction as indicated by a smaller student-faculty ratio and the use of more experienced instructors. Thus, the OLD index was expected to yield a negative correlation with

the measure of student-faculty ratio and a positive correlation with the measure of instructor experience. Although better instruction might be required for more difficult occupations, it also seemed likely that more demanding tasks would also prove more difficult to train. Thus, it was hypothesized that no significant relation would be obtained between OLD scores and behavioral observation ratings of instructional quality. Similarly, the fact that instructional aids, hands-on practice, and feedback might serve to facilitate the learning of more difficult tasks, but might also prove more difficult to develop and apply as occupational difficulty increased, resulted in the expectation that these measures would not yield significant correlations with the OLD index. Finally, despite the fact that yearly flow, manpower requirements, and day length were considered to be significant influences on the training process, trainers' comments that these variables were driven by policy concerns suggested that hypotheses should not be formulated with regard to the OLD index.

Measures of the six student characteristic variables, which Air Force personnel considered to be important descriptors of the training process, are also presented in Table 3. With the exception of academic achievement motivation, measures of all of these variables are obtained as part of the normal selection of assignment process at the time the individual enters the Air Force. Mumford et al. (1987) have indicated that all of these measures have reliabilities above .80. However, note that academic achievement motivation could be measured on the basis of this information. Here the high school and college courses taken by an individual were weighted for their perceived difficulty, as reflected in the average ratings found to have an interrater reliability of .88 by five psychologists. Additionally, it should be recognized that the significance of the age variables lies not in age per se but rather in trainers' belief that, given the average age and typical age range of Air Force enlistees, age provided an approximate index of maturity.

Before concluding this section, we should mention that only one general hypothesis was drawn concerning the relations between these six measures and the OLD index. More specifically, we held that these student characteristics measures would not yield significant relationships with the OLD index. We based this hypothesis on the observation that requirements for admission into specialty training programs are currently set in relation to historic attrition rates and current personnel needs in a specialty field.

The seven training performance variables specified during interviews as being of some importance in describing the training process are presented in Table 3 along with a brief description of the relevant measures. All of these measures were derived from trainees' student evaluation records. These records provide detailed performance descriptions and are maintained with sufficient accuracy to allow some confidence to be placed in their estimates of special individual assistance (SIA) time, washback time, academic counseling, and nonacademic counseling, as well as the reasons specified for these activities. Of course, comparability in the meaning of SIA time, washback time, academic counseling, and nonacademic counseling would be limited under conditions of attrition. Thus, it was necessary to correct estimates of these variables for individuals eliminated from training by increasing the values observed prior to attrition in proportion to the amount of the course that was not completed under the assumption of constant rate. Application of this correction did not seem inappropriate, inasmuch as the split-half reliability coefficients obtained in contrasting the earlier and later phases of the courses yielded a median correlation of .78 among individuals who completed training. The achievement test performance of individuals eliminated from training was measured using their mean score on all tests taken prior to attrition, and it was known that these tests typically yield parallel forms reliability estimates above .80. Finally, note that these corrections were applied only in cases in which elimination occurred because of poor performance attributed by trainers to a lack of adequate ability (academic attrition) or to a lack of adequate motivation

Table 3
Criterion Measures and Hypotheses

Variable	Measure	Expected <i>r</i>	Logic of hypothesis
Course content	Total number of hours of classroom instruction as specified in the 1983 Program of Instruction for the occupational specialty	+	More difficult jobs will take longer time to learn, leading to more hours of classroom instruction
Course length	Total number of unified bodies of instructional material or units as specified in the 1983 Program of Instruction for the occupational specialty	+	More difficult jobs will involve a larger number of more diverse activities to be trained to
Diversity	Average number of instructional hours per unit of material obtained by dividing 1983 course length into 1983 diversity	0	More difficult jobs will be unrelated to practice, due to offsetting effects of course length and task diversity
Practice	Average 5-point rating across five raters of the degree to which course reading materials and 1983 Programs of Instruction emphasized the acquisition of abstract concepts and principles as opposed to specific facts and principles.	+	More difficult jobs will require more intellectual activity and therefore more abstract training materials
Abstract knowledge requirements	Average reading grade level of five paragraphs drawn at random from five 1983 course readings assessed using the Caylor, Stricht, Fox & Ford (1973) algorithm which reflects the number of multisyllable words in passage	+	More difficult jobs will require more advanced reading materials due to greater task complexity
Reading difficulty	Expected attrition rate based on trainers' review of course materials as recorded in the 1983 Program of Instruction for the occupational specialty	+	More difficult jobs will require more difficult training materials due to greater task complexity and diversity, and therefore higher judged attrition
Expected attrition	Number of students per faculty members as specified in the 1983 Program of Instruction for the occupational specialty	-	More difficult jobs will require more intensive supervision of students to ensure adequate task performance
Student-faculty ratio	Average number of months instructors had taught in a training course by June 1983, as indicated in administrative records	+	More difficult jobs will require more experienced instructors due to greater knowledge requirements for effective instruction
Instructor experience	Average of the two most recent supervisory ratings made on an 18-item behavioral observation scale reflecting the quality of instruction provided by all instructors serving in a training program in June of 1983	0	More difficult job training programs may require better instructors but increased difficulty of training materials may make effective instruction more difficult
Instructional quality	Total number of mockup, simulation, and practice devices specified for use in a course in the 1983 Program of Instruction divided by number of hours of classroom instruction as a control variable	0	More difficult job training programs may use aids to offset task difficulty but aids may be more difficult to develop
Instructional aids	Total number of instructional hours devoted to the actual performance of job tasks as specified in the 1983 Program of Instruction divided by number of hours of classroom instruction as a control variable	0	More difficult job training programs may use hands-on practice to offset task difficulty but practice exercises may be more difficult to develop
Hands-on practice	Total number of written tests and quizzes specified in the 1983 Program of Instruction, divided by number of hours of classroom instruction as a control variable	0	More difficult job training programs may give more feedback to offset task difficulty but task difficulty may restrict the amount of feasible feedback over a fixed period of time
Feedback	Total number of students to be trained in 1983 as specified in the Program of Instruction for the job training program	NH	Manpower needs set yearly flow but are not directly relevant to job difficulty
Yearly flow	Availability or nonavailability of a reenlistment bonus in 1983 as specified in Air Force administrative records	NH	Manpower needs set bonus availability but are not directly relevant to job difficulty
Personnel requirements	Whether the course was conducted on an 8-hr (1) or 6-hr (0) instructional day as specified in the 1983 Program of Instruction for the occupational specialty	NH	Day length is often set in relation to local administrative concerns not directly relevant to job difficulty
Day length			

Table 3 (continued)

Variable	Measure	Expected <i>r</i>	Logic of hypothesis
Student characteristics			
Aptitude	Scores on appropriate aptitude area composite of the Armed Services Vocational Aptitude Battery (Weeks, 1984)	0	Requirements for entry based on historic attrition rates and manpower requirements rather than occupational difficulty
Reading level	Total reading grade level derived from scores on the vocabulary and comprehension subtests of the Air Force Reading Ability Test (Mathews & Roach, 1983)	0	Requirements for entry based on historic attrition rates and manpower requirements rather than occupational difficulty
Academic achievement motivation	Number of 42 high school and college prerequisites taken, multiplied by the average of five independent judges' 5-point ratings of course difficulty	0	Requirements for entry based on historic attrition rates and manpower requirements rather than occupational difficulty
Educational level	Level of education at time of enlistment on a 5-point background data scale where 1 indicates no high school degree, 2 indicates a high school degree, 3 indicates some college, 4 indicates a college degree, and 5 indicates some post-graduate work	0	Requirements for entry based on historic attrition rates and manpower requirements rather than occupational difficulty
Educational preparation	Number of high school and college courses taken held to be desirable preparation for entry into a training program according to current Air Force policy	0	Requirements for entry based on historic attrition rates and manpower requirements rather than occupational difficulty
Age	Years of age at time of enlistment held to be an approximate index of maturity in sample at hand	0	Requirements for entry based on historic attrition rates and manpower requirements rather than occupational difficulty
Training performance			
Achievement test performance	Average score on 5 to 10 paper and pencil achievement tests constructed by the measurement and evaluation section on which 70 represents a passing score on a 100-point scale	-	More difficult jobs may be more difficult to learn resulting in poor achievement, but this may be influenced by compensating student inputs and course design
Special individual assistance time	Total number of hours in special individual assistance provided by instructors as a result of poor student performance	+	More difficult jobs will be more difficult to learn and increase the need for special individual assistance and more training time
Academic counseling	Total number of academic counseling sessions conducted in an attempt to remediate poor student performance attributed to academic deficiencies	+	More difficult jobs will be more difficult to learn and increase the need for academic counseling
Nonacademic counseling	Total number of nonacademic counseling sessions conducted in an attempt to remediate poor student performance attributed to lack of motivation	0	Occupational difficulty should not be related to poor performance attributed to motivational problems
Washback time	Total number of retraining hours due to requirements to repeat a section of ■ course after failing to pass quizzes or tests or as a result of counseling	+	More difficult jobs will be more difficult to learn and increase the need for retraining
Academic attrition	Elimination from ■ course on the basis of poor performance attributed to insufficient ability	+	More difficult jobs will be more difficult to learn and increase the number of individuals who can not meet performance expectations
Nonacademic attrition	Elimination from a course on the basis of poor performance attributed to a lack of motivation	0	Occupational difficulty should not be related to poor performance attributed to motivational problems

Note. + = significant positive relationship hypothesized; - = significant negative relationship hypothesized; 0 = no significant relationship hypothesized; NH = no hypothesis generated.

(nonacademic attrition). Thus, cases of death in training or exclusion due to failure to obtain a security clearance were not considered.

The hypotheses formulated concerning the likely relations between these training performance variables and the OLD index are also presented in Table 3. Generally, it was held that negative training outcomes reflecting poor academic performance would increase with the increased difficulty of the tasks one is trained on. Hence, positive correlations were expected between the OLD index and measures of SIA time, washback time, academic counseling, and academic attrition, whereas a negative correlation was expected with the measure of achievement test performance. Because poor performance attributed to motivational deficiencies of the sort found in the military is a function of increasing occupational difficulty, it was anticipated that the OLD index would fail to yield significant relationships with nonacademic counseling and nonacademic attrition. Yet beyond these basic hypotheses, note that training performance is determined by a complex interchange between a number of student characteristics and course content variables. As a result, we assumed that even if the foregoing hypotheses were confirmed, the magnitude of the observed relationships would not be large.

Procedures and Analyses

To examine the construct validity of the OLD index, we obtained information describing the status of each specialty on the various course content variables in a series of site visits to the relevant training centers during the summer of 1984. At this time, the center's registrar's office provided the training evaluation records required for assessing the program's status on the various performance measures. The Air Force Human Resources Laboratory provided all of the information required for assessment of the student characteristics variables. Once these data had been coded in accordance with the format presented earlier, the mean and standard deviation of the student characteristics and training performance measures were obtained within each of the 48 specialties. An attempt was made to control for potential aggregation bias by examining the mean differences among the specialties on these differential variables, and in all cases the resulting *F* ratios were significant beyond the .001 level. Subsequently, the mean and standard deviation of the 48 specialties on the course content, training performance, and student characteristic measures were generated using the specialty field as the unit of analysis. Finally, potential aggregation bias was again examined on the instructor experience and instructional quality measures, and the resulting *F* ratios were sufficient to justify aggregation.

Following these basic descriptive and control analyses, the scores of the specialties on the OLD index were correlated with the scores of the specialties on the course content, student characteristics, and training performance measures. Note that this analysis was carried out on the means of the student characteristics and training performance measures to ensure a level of analysis consistent with interpretation of the OLD index as a measure of occupational, as opposed to individual, characteristics. However, to ensure some generality and control for potential aggregation bias (James, Demaree, & Hater, 1980), similar analyses were conducted at the individual level by assuming that a training program's mean could be applied in describing all members of the relevant occupation. Although it is true that 48 specialties represents a sizable sample of occupations, this sample size is not sufficient to draw stable conclusions concerning the power of the relations obtained between the OLD index and the various criterion variables. Consequently, interpretation of the results obtained at the occupational level necessarily focused on marginal significance levels in relation to the hypotheses specified earlier in the course of this construct validation effort.

Results and Discussion

Table 4 presents the correlations between the various criterion variables and the OLD index at both the individual and

occupational level, and also the mean and standard deviation obtained across the 48 specialties on the OLD index and the student characteristics, training performance, and course content measures. As may be seen, a similar pattern of relations emerged at both the occupational level and the individual level despite the somewhat greater magnitude of the coefficients obtained in the occupational level analyses. Because these results suggested that aggregation bias would not grossly distort any conclusions drawn in the occupational level analyses, and because application of the OLD index is likely to be at this level, the ensuing discussion will focus on the results obtained at the occupational level.

In examining the occupational results, it was found that the OLD index produced significant positive correlations with both the course length ($r = .72, p < .001$) and diversity ($r = .60, p < .001$) measures. These findings are especially noteworthy because they confirm the two most straightforward and unambivalent hypotheses presented. That is, occupational learning difficulty should be closely related to the length of the course and to the amount of material to be mastered. In accordance with expectation, significant positive coefficients were also obtained in relating the OLD index to measures of student-faculty ratio ($r = .78, p < .001$) and instructor experience ($r = .81, p < .001$). Finally, due to the nature of the measures in use and the complex determinants of high-quality instruction, it was not surprising that the practice and instructional quality measures failed to produce significant relations with the OLD index.

On the basis of the fact that the learning difficulty of occupational tasks should be related to the difficulty of training programs, significant positive relationships were hypothesized between the OLD index and measures of abstract knowledge requirements, expected attrition, and reading difficulty. Highly significant positive coefficients were actually obtained for the abstract knowledge requirements ($r = .64, p < .001$) and expected attrition ($r = .64, p < .001$) measures, although a significant coefficient was not obtained for the reading difficulty measure. Although a variety of alternative explanations might account for the failure of the reading difficulty measure to conform to expectation, examination of the raw data suggested that the result might best be attributed to range restriction in the reading difficulty of technical training materials.

The instructional aids, hands-on practice, and feedback measures all failed to produce significant correlations with the OLD index. As we pointed out earlier, these nonsignificant relationships were expected and might be attributed to the trade-off between the need to use these techniques with increasing difficulty and the limitations placed on their use by increasing difficulty. No specific hypotheses were formulated concerning the likely relations between the OLD index and the day length, yearly flow, and personal requirements measures. Relations significant at the .05 level were not obtained for the two former measures, although a highly significant positive relation was achieved between the OLD index and the personnel requirements ($r = .54, p < .001$) measure. In retrospect, it seems that this relationship might be attributed to greater demand in the civilian sector for individuals having training in the more difficult occupational specialties, which leads to higher turnover and greater demand for personnel with this training in the mili-

Table 4
Variable Means, Standard Deviations, and
Correlations With the OLD Index

Variable	M	SD	Level	
			Occupational	Individual
Occupational difficulty				
OLD index	105.410	19.350	—	—
Course content				
Course length	498.720	261.180	.72**	.50
Diversity	72.110	43.420	.60**	.53
Practice	7.690	1.960	-.07	-.18
Abstract knowledge	2.810	0.810	.64**	.46
Reading difficulty	10.910	0.560	.19	.03
Expected attrition	0.095	0.083	.69**	.53
Student-faculty ratio	7.150	3.230	-.75**	-.55
Instructor experience	36.410	12.320	.81**	.59
Instructional quality	2.510	0.120	-.22	-.31
Instructional aids	0.281	0.095	.21	.16
Hands-on practice	0.410	0.126	-.11	-.18
Feedback	0.350	0.120	-.07	-.07
Yearly flow	741.410	715.610	-.20	-.20
Personnel requirements	0.340	0.470	.54**	.42
Student characteristics				
Aptitude composite				
score	72.560	7.240	.16	.05
Reading grade level	11.510	0.310	.24	.06
Academic achievement				
motivation	39.950	4.130	.33*	.07
Educational level	2.150	0.080	-.23	-.01
Specific educational				
preparation	1.520	0.400	-.01	-.11
Age	20.200	2.830	.01	.03
Training performance				
Achievement test				
performance	86.520	2.750	.32*	.08
Special individual				
assistance time	8.190	7.110	.27*	.05
Academic counseling	1.380	0.850	.25*	.05
Nonacademic				
counseling	0.100	0.810	.00	.00
Washback time	12.120	8.470	.33*	.06
Academic attrition	0.021	0.021	.13	.02
Nonacademic attrition	0.003	0.006	.05	.00
Sample size	—	—	48	5,970

Note. Significance levels are determined only for occupational level results due to extremely large sample size in individual level analyses. OLD = occupational learning difficulty.

* $p < .05$. ** $p < .001$.

tary. Moreover, it is possible that the insignificant relation obtained for yearly flow might be due to the need for large numbers of individuals in occupations of varying difficulty.

Based on the fact that Air Force placement policy is driven by training performance and manpower requirements, it was not expected that significant correlations would be obtained between the OLD index and the measures of student characteristics. Overall, this expectation was borne out, inasmuch as the aptitude, reading grade level, educational level, specific educational preparation, and age measures failed to yield correlations with the OLD index significant at the .05 level. However, the

academic achievement motivation measure did produce a significant positive relation with the OLD index ($r = .33, p < .05$). Although this relationship reached only marginal significance levels, it was not expected and may reflect the tendency of more difficult occupational specialties to require greater academic motivation in obtaining the requisite preparation.

A more complex pattern of results emerged among the training performance measures. In accordance with the hypothesis that negative training outcomes should occur more frequently in training for more difficult occupational specialties, significant positive correlations were observed between the OLD index and measures of SIA time ($r = .27, p < .05$), academic counseling ($r = .21, p < .05$), and washback time ($r = .33, p < .05$). Moreover, because nonacademic counseling and nonacademic attrition focus on motivational issues, we held that these measures would not be correlated with the OLD index. Correspondingly, both of these measures failed to produce correlation coefficients significant at the .05 level. Yet, this apparent confirmation of the foregoing hypothesis must be approached with some caution. The academic attrition measure failed to produce the expected positive relation with the OLD index, perhaps due to the low frequency of academic attritions. Thus, it is possible that the failure of the nonacademic attrition measure to yield a significant correlation may also be attributed to the limited number of nonacademic attritions observed in the sample.

Although a significant negative correlation was expected for achievement test performance, this measure produced a marginally significant positive relation with the OLD ($r = .32, p < .05$) index. This result might be attributed to increases in training program length, instructor availability, and instructor experience with increasing difficulty, all of which should generate better student performance on standardized achievement tests. Note that some support for this hypothesis has been obtained in a separate study conducted by Mumford et al. (1987), which also suggests that variation in student characteristics might in part account for this result.

Obviously, the foregoing results did not confirm all of the hypotheses generated over the course of this construct validation effort. However, of the 25 hypotheses proposed for evaluating the construct validity of the OLD index, only 4 were disconfirmed. Moreover, the OLD index produced the significant positive relations with measures of course length, diversity, expected attrition, and abstract knowledge requirements that would be expected of any index that attempted to capture the learning difficulty of tasks performed in entry-level positions within an occupational specialty. Finally, it should be recognized that in cases in which hypotheses could not be formulated a priori and in which significant relationships emerged, these relations proved to be interpretable, given the purported meaning of the OLD index and the nature of the training system under consideration.

Beyond these basic findings, one other important piece of evidence pointed to the construct validity of the OLD index. We refer to the apparent ability of the OLD index to display some convergent and discriminant validity. More specifically, it was found that this measure of task learning time was related to negative academic or learning deficiency outcomes, but not to negative motivational outcomes. Furthermore, in accordance

with expectation, the OLD index yielded significant relations with course content but not with student characteristic variables. Although certain ambiguities remain with respect to the interpretation of these correlational findings, it does seem that the overall pattern points to the convergent and discriminant validity of the OLD index as an aggregate measure of the learning difficulty of occupational tasks.

Despite the construct validation evidence already discussed, it should be recognized that certain ambiguities are associated with interpretation of the OLD index as a measure of occupational difficulty. First, it is possible that common occupational stereotypes might be driving the correlational pattern described. Although the explanation can not be completely ruled out, given the evidence cited earlier and the weak relationships observed between the OLD index and the student characteristic variables, it is not a plausible one. Second, it might be argued the judge's knowledge of current course length determined their evaluations of task learning time. Of course, this hypothesis implies that the same task will be given different ratings by judges drawn from occupational fields having different course lengths and that judges within an occupational field would rate all tasks in a similar fashion. Thus, an argument contradicting this hypothesis may be found in Burtch et al.'s (1982) observations, indicating that there is substantial variability in average learning time ratings obtained within an occupational field. Moreover, their finding that judges display substantial agreement in their learning time evaluations of tasks drawn from 15 specialties that have markedly different course lengths also, argues against this hypothesis. Further evidence in this regard may be obtained from the path analysis conducted by Mumford et al. (1987), which indicates that occupational difficulty, as measured by OLD values, is a cause of course length, and from the support for the construct validity of the OLD index obtained, using training performance variables that should not be susceptible to this effect.

If one grants the appropriateness of the foregoing arguments, then the results obtained in the present effort appear to provide strong support for the construct validity of the OLD index. More generally, they point to the feasibility of examining the construct validity of job analysis measures through careful definition of the relevant job characteristics, systematic measurement of these characteristics, and assessment of the relation of the measures to other relevant constructs in a well-defined nomological net. Although such systematic construct validation efforts have frequently been conducted in the areas of personnel selection, training evaluation, and performance appraisal (Goldstein, 1974; Guion, 1965; Latham & Wexley, 1981), relatively little effort has been expended on studies concerned with the construct validity of job analysis measures. This trend appears to result from an abiding belief that a direct and comprehensive assessment of job activities will yield measures of obvious content validity (Fine, personal communication, May 15, 1984). Content validity, however, is only one source of evidence pertaining to the construct validity of job analysis measures, and exclusive reliance on such evidence may occasionally yield misleading conclusions (Wagner, 1985). Moreover, reports by Chesler (1948), Rupe (1956), Fleishman and Quaintance (1984), and Fleishman and Mumford (1987) have all shown that systematic construct validation efforts of

the sort previously described may do much to enhance understanding of the relative strengths and weakness of alternative job analysis measures, thus providing a firmer groundwork for the description of job characteristics.

Certainly, only limited confidence could be placed in the practical value of the OLD index without such evidence pertaining to its construct validity. Yet, given this evidence, it does not seem unreasonable to conclude that the OLD index might prove of value in addressing a variety of practical problems. For instance, in many settings, trainers are required to define course length and allocate time to training in alternative job tasks when firm quantitative data providing guidelines for course structure is not available. Here OLD information might be used to define an appropriate course length while specifying the relative amount of time spent in training for various tasks. With regard to manpower allocation, OLD values might be used to specify jobs for which experienced help should be sought, as well as jobs that require more talented trainees due to the high learning difficulty of job tasks. In a related vein, occupational difficulty has been a perennial concern in job evaluation efforts (McCormick, 1979). Thus, the evaluations of task learning time provided by the OLD index might be used in attempts to define compensation rates and comparable worth.

When these varied applications of the OLD index are considered in light of construct validation strategy used, they lead to one further conclusion. Traditionally, job analysis efforts have tended to focus on the comprehensive description of job activities rather than on the identification and assessment of specific activity characteristics of value in formulating personnel policy (Fleishman & Mumford, 1987). Although these comprehensive activity descriptions may have substantial value for certain purposes, they rarely provide unambiguous guidance for personnel decisions, due to the breadth and complexity of this descriptive information (Fleishman & Mumford, 1987; Fleishman & Quaintance, 1984). We hope that the present study has shown that it is possible to develop psychologically meaningful constructs and measures capable of summarizing this mass of descriptive data in a manner that facilitates valid personnel decisions for specific administrative purposes.

Given the apparent value of this approach in the case of the OLD index, it would seem that efforts examining the meaningfulness of constructs such as discretionary task performance (Jaques, 1977; Korotkin, Mumford, Levin, Wallis, & Fleishman, 1985), information acquisition requirements (Mumford, 1986), and work load (Chiles, 1982) might also prove valuable. If further efforts along these lines are coupled with systematic construct validation studies, they might do much to provide a more sophisticated basis for describing and understanding the relation between job activities and human performance.

References

- Burtch, L. D., Lipscomb, M. S., & Wissman, D. J. (1982). *Aptitude requirements based on task difficulty: Methodology for evaluation* (AFHRL-TR-81-34). Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.
- Caylor, J. S., Sticht, T. G., Fox, L. C., & Ford, J. P. (1973). *Readability of job materials*. Alexandria, VA: Human Resources Research Organization.

- Chesler, D. J. (1948). Reliability and comparability of different job evaluation systems. *Journal of Applied Psychology*, 32, 465-475.
- Chiles, D. W. (1982). Workload, task, and situational factors as modifiers of complex human performance. In E. A. Alluisi & E. A. Fleishman (Eds.), *Human performance and productivity: Stress and performance effectiveness* (pp. 11-56). Hillsdale, NJ: Erlbaum.
- Christal, R. E. (1974). *The United States Air Force Occupational Research Project* (AFHRL-TR-73-75). Lackland AFB, TX: Air Force Human Resources Laboratory, Occupational Research Division.
- Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods*. New York: Irvington.
- Fleishman, E. A., & Mumford, M. D. (1987). The ability requirements scales. In S. Gael (Ed.), *Job analysis handbook*. New York: Wiley.
- Fleishman, E. A., & Quaintance, M. K. (1984). *Taxonomies of human performance: The description of human tasks*. Orlando, FL: Academic Press.
- Fugill, J. W. K. (1972). Task difficulty and task aptitude benchmark scales: A feasibility study in mechanical, electronic, administrative, and general job areas. *Proceedings of 4th Annual Conference of the Military Testing Association*, 4, 301-304.
- Gettinger, M., & White, M. A. (1979). Which is the stronger correlate of school learning? Time to learn or measured intelligence? *Journal of Educational Psychology*, 71, 405-412.
- Goldstein, I. L. (1974). *Training: Program development and evaluation*. Monterey, CA: Brooks/Cole.
- Guion, R. M. (1965). *Personnel testing*. New York: McGraw-Hill.
- James, L. R., Demaree, R. G., & Hater, J. J. (1980). A statistical rationale for relating situational variables and individual differences. *Organizational Behavior and Human Performance*, 25, 354-364.
- Jaques, E. (1977). *A general theory of bureaucracy*. London: Heineman.
- Korotkin, A. L., Mumford, M. D., Levin, K. Y., Wallis, R. M., & Fleishman, E. A. (1985). *Taxonomic efforts in the description of leadership behavior: A general approach*. Bethesda, MD: Advanced Research Resources Organization.
- Krumboltz, J. D. (1965). *Learning and the educational process*. Chicago: Rand McNally.
- Latham, G. P., & Wexley, K. N. (1981). *Increasing productivity through performance appraisal*. Reading, MA: Addison-Wesley.
- Leczmar, W. B. (1971, July). *Three methods for estimating the difficulty of job tasks* (AFHRL-TR-71-30). Lackland AFB, TX: Air Force Human Resources Laboratory, Personnel Division.
- Madden, J. M. (1962). What makes work difficult? *Personnel Journal*, 41, 341-344.
- Mathews, J. J., & Roach, B. W. (1982). *Reading abilities tests: Development and norming for Air Force use* (AFHRL-TR-82-26). Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.
- McCormick, E. J. (1979). *Job analysis: Methods and applications*. New York: Amacom.
- Mead, D. F., & Christal, R. E. (1970). *Development of a constant standard weight equation for evaluating job difficulty* (AFHRL-TR-70-44). Lackland AFB, TX: Air Force Human Resources Laboratory, Personnel Division.
- Mumford, M. D. (1986). Leadership in the organizational context: A conceptual approach and its applications. *Journal of Applied Social Psychology*, 16, 508-531.
- Mumford, M. D., Harding, F. D., Fleishman, E. A., & Weeks, J. L. (1987). *An empirical model for use in assessing the impact of aptitude requirement adjustments on Air Force resident technical training* (AFHRL-TR-86-19). Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.
- Rupe, J. C. (1956). *Research into basic methods and techniques of Air Force job analysis, IV* (AFPTRC-TN-56-51). Chanute AFB, IL: Air Force Personnel and Training Research Center, Air Research and Development Command.
- Wagner, M. (1985, August). *On the use of maintenance reports in job analysis*. Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles.
- Weeks, J. (1984). *Occupational learning difficulty: A standard for determining the order of aptitude requirement minimums* (AFHRL-SR-84-26). Brooks AFB, TX: Air Force Human Resources Laboratory.

Received December 23, 1985

Revision received April 13, 1987

Accepted May 4, 1987 ■

Is Cost Accounting the Answer? Comparison of Two Behaviorally Based Methods for Estimating the Standard Deviation of Job Performance in Dollars With a Cost-Accounting-Based Approach

Olen L. Greer

Department of Accounting
Southwest Missouri State University

Wayne F. Cascio

Graduate School of Business and Administration
University of Colorado, Denver

Accurate estimation of the standard deviation of job performance in dollars (SD_y) can improve the precision of utility estimates of expected payoffs from personnel programs. The purpose of this study was to compare directly the estimates of SD_y obtained using a cost-accounting-based estimate of SD_y , the Global Estimation Model, and the CREPID procedure. The study was conducted in a large, soft-drink bottling company. Each method for estimating SD_y was applied to the job classification, *route salesman*, producing three independent estimates of SD_y . These estimates were tested for significant differences. Results indicated that the Global Estimation Model estimate and the cost-accounting-based estimate were not significantly different, whereas the estimate produced by the CREPID procedure was significantly smaller. Limitations of the cost-accounting-based estimate are identified and results are discussed in terms of their implications for the theory and practice of utility analysis in organizations.

Utility analysis (or decision theory) is the determination of institutional gain or loss anticipated to result from various courses of action (Cascio, 1987a). The utility of a selection device is the degree to which its use improves the quality of the individuals selected, beyond what would have occurred had that device not been used (Blum & Naylor, 1968). A number of utility models have been proposed; the best known are those of Taylor and Russell (1939), Naylor and Shine (1965), Brogden (1949), and Cronbach and Gleser (1965). The Brogden–Cronbach–Gleser model is of particular interest, as it is the only model that incorporates the concept of cost of selection, or dollars gained or lost, into the utility index (Cascio, 1987b). This utility approach is intuitively appealing because outcomes are expressed in a form comprehended and appreciated at all levels of management: dollars. The Brogden–Cronbach–Gleser model expresses the utility index in the form of an increase (decrease) in the dollar payoff of the selected group.

Of the data requirements of this utility model, the standard deviation of job performance, SD_y , is the most difficult parameter to estimate. The difficulties associated with producing accurate estimates of SD_y have long been the major impediment to the widespread implementation of the Brogden–Cronbach–Gleser model by personnel psychologists (Bobko, Karren, & Parkington, 1983; Burke & Frederick, 1984; Cascio, 1980; Schmidt, Hunter, McKenzie, & Muldrow, 1979; Schmidt, Mack, & Hunter, 1984; Weekley, Frank, O'Connor, & Peters, 1985). Three major approaches to the estimation of SD_y have been advanced: the Global Estimation Model (Schmidt et al.,

1979), the Cascio–Ramos estimate of performance in dollars (CREPID; Cascio, 1987b; Cascio & Ramos, 1986), and the cost-accounting approach (Roche, 1961). The first two approaches are behaviorally based, whereas Roche's approach was based on the use of cost-accounting data. Note that Roche was a personnel psychologist working in an accounting arena, which resulted in his heavy reliance on the input of the firm's accountants (Cronbach & Gleser, 1965). This introduces the possibility of contaminants, in the sense that a nonaccountant was working with data of which he had relatively little understanding. In addition, it may be (as Cronbach & Gleser pointed out) that the organization's accountants did not clearly understand the problem, further complicating matters.

In spite of these and other criticisms of the study (e.g., Boudreau, 1983), a cost-accounting approach to the estimation of SD_y remains as the conceptual standard of comparison (Cascio, 1987b), due to the fact that accounting data generally are objective, verifiable by a third-party observer, and subject to both internal and external audit (Bedford et al., 1966; Boatsman et al., 1977). However, the use of cost-accounting data to estimate SD_y is both costly and time consuming (Cronbach & Gleser, 1965), inasmuch as the typical cost object is a unit of product, not human performance levels (Cherrington, Hubbard, & Luthy, 1985; Horngren, 1982). The cost object in Roche's (1961) study was an individual worker's performance level. For a cost-accounting system to provide a valid estimate of the cost of any cost object, it should be designed with that cost objective in mind. The cost objective and cost object were significantly different from the cost objectives and cost objects that the cost-accounting system was designed to accommodate. This led to many assumptions, estimates, and arbitrary allocations (Cronbach & Gleser, 1965).

Traditionally, cost-accounting systems have not established the individual's worth as a cost object, although in the accounting field, human resource accounting (HRA) represents an at-

This study was conducted as part of a doctoral dissertation by the first author, under the chairmanship of the second author.

Correspondence concerning this article should be addressed to Olen L. Greer, Department of Accounting, Southwest Missouri State University, 901 South National Avenue, Springfield, Missouri 65802.

tempt to do so. A review of the various HRA models is beyond the scope of this article; however, it is sufficient to point out that, at present, the HRA movement has run aground due to the difficulties associated with operationally defining a relatively soft concept: the value of the human worker (DeAngelo, 1982; Dittman, Juris, & Revsine, 1976, 1980). Thus, HRA research has failed to provide either an acceptable method for valuing the human asset as a balance-sheet item or for calculating or estimating the value of SD_y . As a result, the accounting systems of organizations remain ill equipped to provide cost data when worth of the human asset is involved. In fact, behavioral methods are more feasible to implement in business settings (Weekley et al., 1985) because (a) the methodology required to estimate SD_y , using either the Global Estimation or the CREPID models, is specified clearly, (b) these procedures can be applied without regard to the nature of the business (profit or nonprofit; service, merchandising, or manufacturing; etc.), and (c) these procedures can be applied without regard to the type of accounting or management information system used by the firm (e.g., standard cost system or normal cost system; computerized or manual; Reilly & Smither, 1985).

However, a fundamental concern related to both of these methodologies is the accuracy or validity of the estimates of SD_y obtained (Reilly & Smither, 1985; Weekley et al., 1985). Numerous studies have examined and compared the results of alternative procedures for estimating SD_y (Bobko et al., 1983; Burke & Frederick, 1984; Landy, Farr, & Jacobs, 1982; Reilly & Smither, 1985; Weekley et al., 1985). Bobko et al. (1983) compared estimates of SD_y derived from using the Global Estimation Model (Schmidt et al., 1979) with actual standard deviations of performance measures. They found that the estimates were not statistically different from (and, in fact, were quite close to) the actual standard deviation. In addition, they found strong support for the assumption of underlying normality for the objective measure of job performance. Finally, they found that although supervisors tend to underestimate actual performance values, the effect of the underestimation was reduced substantially when the difference in percentiles (the 85th percentile minus the 50th percentile and the 50th percentile minus the 15th percentile) was computed to estimate SD_y . In summary, the Bobko et al. study provided evidence that estimation of SD_y is not necessarily the Achilles' heel of utility analysis.

Two other studies (Reilly & Smither, 1985; Weekley et al., 1985) compared the Global Estimation Model and CREPID and concluded that they produce different results. Both sets of authors called for comparative research in a field setting that also includes meaningful external criteria such as cost-accounting outcomes.

The purpose of this study was to compare directly the estimates of SD_y using a cost-accounting-based estimate of SD_y (SD_{y1}), the Global Estimation Model (SD_{y2}), and the CREPID procedure (SD_{y3}). The cost-accounting-based estimate was used as the standard of comparison for the other methodologies. The specific null hypotheses tested were

$$H_{01}: SD_{y1} = SD_{y2},$$

$$H_{02}: SD_{y1} = SD_{y3},$$

and

$$H_{03}: SD_{y2} = SD_{y3}.$$

Method

Subjects

The major consideration in the design of a study of this nature is the identification and selection of a suitable data base, that is, the identification of an organization that meets the following criteria:

1. Cooperation—is willing to make the cost-accounting records and supervisory personnel available to the researcher.
2. Size—has a job classification in which there are significant numbers of employees to ensure adequate statistical power.
3. Performance—variability in performance levels must make a difference in the selected job classification; that is, high performance levels should be distinguishable from average performance levels in dollars. The use of utility analysis presupposes that differences in performance levels are important to the organization (Cronbach & Gleser, 1965).

The study was conducted in a midwestern soft-drink bottling company that manufactures, merchandises, and distributes nationally known products. The organization employs 221 people and serves a 25,000-sq mi region. Data were provided by 29 supervisors ($N_1 = 29$) and from the accounting records of the firm. The average management experience of the supervisors was 9 years, 3 months (range = 9 months–30 years, 6 months). All of the supervisors were White men, and their average age was 40 years, 6 months (range = 29 years–60 years).

The job classification used was *route salesman*¹ ($N_2 = 62$). All route salesmen were White men. The average age of the route salesmen was 32 years, 2 months (range = 21 years, 1 month–53 years, 1 month). The average number of years of sales experience of the route salesmen was 7 years, 5 months (range = 9 months–34 years, 6 months). This job classification was selected for two reasons: (a) There were a large number of individuals in this job classification, and (b) variability in performance levels had a direct impact on output. Therefore, differences in job performance created significant differences both in payoff for the company and for the individual route salesman. Route salesmen were paid on the basis of a small weekly base wage and a commission per case of soft drink sold.

Procedures

Global Estimation Model. Following the questionnaire-based procedures developed by Schmidt et al. (1979), supervisors were asked to estimate the dollar value of route salesmen's job performance at the 15th, 50th, and 85th percentiles of the performance distribution. Differences between estimates of performance at the 15th and 50th percentiles, as well as between the 50th and 85th percentiles, were computed for each participant. The mean dollar value difference between low and average performance and the mean dollar value difference between average and high performance were then computed. These difference scores each represent estimates of SD_y , under the assumption that job performance outcomes are distributed normally. However, the final Global Estimation Model estimate of SD_y was obtained by averaging these separate SD_y estimates. All of the supervisors provided estimates of dollar value that were consistent with percentile magnitudes; that is, performance at the 50th percentile level was valued higher than performance at the 15th percentile level, and so forth.

CREPID. This procedure involved eight steps, as described by Cascio and Ramos (1986). Each route salesman was rated by two raters, a primary rater and a secondary rater, in order to assess interrater reliability.

Cost-accounting method. The cost-accounting method involved the following steps:

¹ The authors are acutely aware of the importance of using nonsexist language. However, as all of the subjects in the study were men, the most accurate term was *route salesman*.

1. Output data on each of the route salesmen were collected from the records of the organization for a 1-year period. This was done to eliminate seasonality from the data. These output data were expressed in terms of number of cases sold and by package size and type.

2. The weighted average sales price per case of product (SP_u) by package size and package type was calculated using data provided by the accounting department. In calculating the SP_u , discounts were taken into consideration.

3. The variable cost per unit (VC_u) by package size and package type was calculated using data provided by the accounting department. Variable costs are costs that vary in total as the volume of production or sales changes (Cherrington et al., 1985). Note that on a per-unit basis, variable costs are fixed (Cherrington et al., 1985). The VC_u was composed of direct labor, direct materials (syrup cost, CO₂ gas, crowns, closures, and bottles), variable factory overhead (state inspection fees, variable indirect materials, and variable indirect labor), and variable selling expenses (route salesman's commission).

4. Using the information produced in the preceding steps, 2 and 3, the contribution margins per case of product (CM_u) by package size and type were calculated. Contribution margin per unit is defined as sales price per unit minus the variable cost per unit (Horngren, 1982). Fixed costs were not included in the calculation of the contribution margins. Fixed costs are costs that remain constant in dollar amount as volume of production or sales changes (Cherrington et al., 1985). Examples of fixed costs include fixed factory overhead (depreciation on plant equipment, factory supervision, etc.), general and administrative expenses (corporate officers salaries, clerical wages, office supplies, etc.), and fixed selling expenses (salaries of sales supervisors, advertising, etc.).

5. The contribution margins calculated in Step 4 were multiplied by the output figures (Step 1), thus producing a total contribution margin for the route salesman for the year. This figure represents the total amount (in dollars and cents) contributed toward the coverage of fixed costs, and then profit, by each route salesman.

6. The percentage of the total contribution margin attributable to the route salesman was calculated on a route-by-route basis. To accomplish this, the sales of each route were partitioned into two categories, *home market* and *cold bottle*. Data were available with respect to the specific percentages of sales of the respective routes that were home market and cold bottle. Home market represents sales in large supermarkets and chain stores, in which the product is purchased and taken home to consume. Cold bottle represents sales in small convenience stores, filling stations, restaurants, and vending operations, in which the product generally is purchased and consumed on location. According to a consensus of top management, home market sales are influenced less by the efforts of the route salesman than by such variables as national advertising, the goodwill associated with the product itself, and the efforts of the route salesman's supervisor. On the other hand, the route salesman exercises greater influence over the relative sales level in the cold-bottle market, due to the fact that the supervisor makes few calls on the customer, and the route salesman has a greater degree of flexibility in offering price incentives, seeking additional display space, and so forth. The critical question was, "How much influence does the route salesman have in each of the respective sectors?" An attempt was made to isolate statistically the effect on output of performance or effort by the route salesman, using stepwise regression. However, measures of the variables that affect output on a given route were not available, such as effort or activity of the route salesman's supervisor, the impact of national and local advertising, and activity of major competitors within a route. Therefore, the percentage of sales or contributions attributable to the efforts of the route salesman was determined by a consensus of six top managers, facilitated by a Delphi technique.² The "problem" confronting the management of the organization in the current study was to determine the percentage of contribution attributable to the efforts of route salesmen within the home market and cold bottle sectors. The following instructions were communicated to the six top managers within the organization:

Table 1
Percentage Estimates of the Influence of Route Salesmen in Home and Cold-Bottle Markets Using the Delphi Technique

Manager	Estimate 1		Estimate 2		Estimate 3	
	Home market	Cold bottle	Home market	Cold bottle	Home market	Cold bottle
1	50	90	40	75	20	30
2	20	30	20	30	20	30
3	20	20	20	30	20	30
4	25	90	20	30	20	30
5	25	70	25	60	20	30
6	20	20	20	30	20	30
M	26.7	53.3	24.2	45.0	20.0	30.0
SD	11.7	32.2	6.3	19.9	0	0

While it is recognized that the Route Salesman himself can affect significantly the level of sales output of his route, there are other variables that affect output as well. Such variables include (but are not limited to) the sales-related efforts of the Route Salesman's supervisor, advertising (both national and local), the performance of competitors within the Route Salesman's territory, the route composition (i.e., the number of large, supermarket stores versus small, "Mom and Pop" operations), the temperature and general weather conditions, and the goodwill associated with the product itself. In this study, we are trying to isolate the effect of performance of the Route Salesman. That is, "What percentage of total output would you attribute to the efforts of a Route Salesman within the "Home Market" and "Cold Bottle" sectors, respectively?" We realize that this is a difficult judgment to make, but take into account two factors:

1. You will be provided with the estimates of other supervisors along with the rationale behind their decisions.
2. You will be allowed to adjust your estimate, after taking into account the additional information.

Please complete the following sentence: In my opinion, the efforts of a Route Salesman account for _____ percent of the total output within 'Home Markets' and _____ percent of the total output within 'Cold Bottle' markets. The rationale behind these estimates is _____.

Convergence was obtained after three sets of independent estimates by the six judges. Table 1 summarizes their responses. As Table 1 demonstrates, the top management of the organization placed the portions of home market sales and cold-bottle sales attributable to the efforts of the route salesman at 20% and 30% respectively.

7. The percentages calculated in Step 6 were multiplied by the respective total contribution margins calculated in Step 5, yielding the total contribution margin attributable to each route salesman. An example calculation is shown in Table 2. This amount served as the cost-accounting-based estimate of each route salesman's worth to the organization.

8. The mean and standard deviation of the preceding values were calculated. This standard deviation served as the cost-accounting-based estimate of SD_y .

This approach can be characterized as a *contribution* approach to costing the performance of the route salesmen. It differs from Roche's (1961) approach in that Roche used "the profit attributable to each Radial Drill Operator" as the surrogate measure of worth. The advan-

² The Delphi technique was developed by researchers at the Rand Corporation to interchange employee's ideas and feedback while avoiding the inefficiencies and inhibitions of face-to-face groups (Dalkey & Helmer, 1963).

Table 2
Sample of the Total Contribution Attributable to Route Salesman A (RSA) Using Cost-Accounting Procedures

Product	SP_u	—	VC_u	=	CM_u	×	Sales output ^a	=	GCM
1	5.00	—	2.75	=	2.25	×	40,000	=	90,000
2	7.60	—	4.85	=	2.75	×	20,000	=	55,000
3	8.30	—	5.65	=	2.65	×	15,000	=	39,750
Gross contribution generated by RSA									= 184,750

Note. If RSA sells 45% of his output in home markets ($HM\% = 0.45$) and 55% in cold-bottle markets ($CB\% = 0.55$), and given the influence ratios (IR_{hm} , IR_{cb}) of 20% and 30%, respectively (determined by the Delphi technique), then the total contribution attributable to RSA ($TOTCM_{RSA}$) is calculated as follows:

$$TOTCM_{RSA} = GCM \times (IR_{hm} \times HM\%) + (IR_{cb} \times CB\%)$$

$$TOTCM_{RSA} = \$184,750 \times (.20 \times .45) + (.30 \times .55)$$

$$TOTCM_{RSA} = \$47,111.25.$$

SP_u = sales price per unit of product; VC_u = variable cost per unit of product; CM_u = contribution margin per unit of product; GCM = gross contribution margin. Figures, except ^a, are in dollars.

^a Numbers of cases.

tage of the contribution approach lies predominantly in the fact that there has been no arbitrary allocation of fixed costs. Boudreau (1983) pointed out that Roche's inclusion of fixed costs rendered an underestimation of selection utility, due to the fact that fixed costs are essentially a cost of being in business. Allocation of these costs to employees is inappropriate because personnel programs generally will not change, or impact on, these costs. Although contribution costing is not generally accepted for external reporting purposes, it is generally accepted and, in fact, recommended for internal, managerial reporting purposes (Cherrington et al., 1985; Horngren, 1982).

Results

The cost-accounting-based procedure produced an estimate of SD_y (SD_{y1}) of \$15,864, with a mean value of job performance of \$44,985. Estimates of worth ranged from \$11,237 to \$99,557. These values were skewed positively ($Q_3 - Q_2 = \$12,175$, greater than $Q_2 - Q_1 = \$5,282$). This result was anticipated, inasmuch as the values were calculated for experienced job incumbents. Assuming that the selection system of the organization is valid, the expected result would be positive skewness, as the lower level job performers already would be eliminated. This phenomenon will be discussed in more detail in a later section of this study. A chi-square goodness-of-fit test was performed to test the normality of the distribution. The result was failure to reject the null hypothesis that the distribution is normally distributed, $\chi^2(5, N = 62) = 7.10, p > .025, ns$.

The wide range of experience levels (6 months–34 years, 6 months) of route salesmen was a matter of concern. That is, one might hypothesize that the variability in output as expressed in the contribution values used in the cost-accounting-based approach can be attributed to seniority and experience levels. The

correlation between amount of experience as a route salesman and the cost-accounting-based estimates of worth was calculated. The resulting correlation coefficient ($r = 0.118, p > .10, ns$) indicated that little, if any, correlation exists between the two variables. This result corresponded to the predictions and comments made by company management.

The Global Estimation Model estimate of SD_y (SD_{y2}) was \$14,636, calculated by averaging the differences between the 15th and 50th percentiles (\$14,834) and the 50th and 85th percentiles (\$14,439). Estimates of worth at the 15th percentile ranged from 0 to \$80,000 ($M = \$17,145$; $SD = \$14,633$). Comparable figures for the 50th and 85th percentiles were, respectively, \$10,000 to \$120,000 ($M = \$31,979$; $SD = \$24,829$) and \$15,000 to \$175,000 ($M = \$46,417$; $SD = \$38,443$). The estimate of SD_y obtained by differencing the 15th and 50th percentile values was compared with the estimate of SD_y obtained by differencing the 85th and 50th percentiles, and no significant difference was detected, $t(56) = 0.085, p > .10, ns$. Because there was no statistically significant difference between these means, the assumption of normality was supported. This result corresponds to the findings of Bobko et al. (1983), discussed earlier.

To evaluate differences in results produced by the cost-accounting procedure and the Global Estimation Model, 90% confidence intervals were constructed. If the intervals overlap, then we can conclude (with 90% confidence) that results from these two procedures do not differ significantly. The upper and lower limits of the 90% confidence interval for the cost-accounting model were \$18,208 and \$13,520, respectively. The upper and lower limits of the Global Estimation Model were \$19,780 and \$9,492, respectively. Because there was overlap between the confidence intervals derived from the two methods, we concluded that these estimates were not significantly different from each other.

The CREPID estimate of SD_y (SD_{y3}) was \$8,988, with a mean value of job performance of \$38,435. Estimates of worth ranged from \$19,890 to \$53,171. As would be expected when dealing with experienced job incumbents, the data were skewed positively, but not to a degree that would jeopardize the underlying assumptions of the t test. A chi-square goodness-of-fit test resulted in failure to reject the null hypothesis that the data are normally distributed, $\chi^2(5, N = 62) = 11.51, p > .025, ns$. Moderate departures from the assumption of equal variances have been shown to have negligible effect on the operating characteristics of the t ratio (Boneau, 1960; Baker, Hardyck, & Petrinovich, 1966; Hardyck & Petrinovich, 1969).

In a previous article (Cascio & Ramos, 1986), the authors noted that although a modified magnitude estimation procedure was used in the CREPID performance scale, the underlying distribution is based on a percentile distribution that is rectangular in shape. A rectangular distribution has a fixed mean and standard deviation. The CREPID ratings would be expected to demonstrate a fixed mean and standard deviation only in the long run for each individual rater, inasmuch as each rater's definition of performance at the 25th, 50th, 75th, and so on, percentiles is likely to vary. When performance ratings are pooled across raters, as they were in this study, the distribution will not be uniform in shape, but rather will be skewed or normal. The use of such a distribution actually facilitates subse-

Table 3
Standard Deviation Estimates Using Three Methods

Method	SD _y estimate	M worth	M value	Range	Test of significance		
					SD _{y2}	SD _{y3}	SD _{y1}
Cost accounting (SD _{y1})	\$15,864 (100%)	\$44,985 (100%)	\$41,166 (100%)	\$11,237–\$99,557	90% CI (ns)	—	—
Global estimation model (SD _{y2})	\$14,636 (92%)	\$31,979 (71%)	\$23,000 (56%)	0–\$175,000	—	90% CI (ns)	—
CREPID (SD _{y3})	\$8,988 (57%)	\$38,435 (85%)	\$40,489 (98%)	\$16,855–\$53,171	—	—	<i>t</i> (60) = 5.344, <i>p</i> < .10

Note. The percentage figures after each of the dollar values represent the dollar values expressed as a percentage of the corresponding cost-accounting-based value. SD_y = standard deviation of job performance; CI = confidence interval; CREPID = Cascio–Ramos estimate of performance in dollars.

quent data analyses because it permits a researcher to use parametric statistics.

As noted earlier, two sets of performance ratings were obtained on each route salesman, one provided by the route salesman’s first-level supervisor, and a corresponding set provided by the route salesman’s second-level supervisor. As a method for testing the degree of interrater agreement, 26 two-way analyses of variance (ANOVAs) were performed, one for each individual principal activity.³ The dependent variable was the appraisal rating assigned by each of the two individual raters to the ratees. The two independent variables were *rater* (two levels, primary and secondary) and *ratee* (*n* = 62). Using the results of these ANOVAs, we computed coefficients alpha for each of the 26 principal activities. These coefficients ranged from 0.32 to 0.83, with a mean value of 0.62, computed using a Fisher *r*-to-*z* transformation. The interrater reliability of the estimates of overall worth assigned to ratees using the CREPID procedure was 0.70.

One would expect the cost-accounting-based estimate and the CREPID procedure to be correlated, as both estimates were derived from the same group of subjects. In fact, the correlation was *r* = 0.118, *p* < .10. When the data sets are correlated, the appropriate test of significance is a *t* test (McNemar, 1969).⁴ Applying this test to the second null hypothesis (*SD*_{y1} = *SD*_{y3}) resulted in rejection of the null hypothesis, *t*(61) = 5.344, *p* < .10. For the CREPID ratings, the lower and upper limits of a 90% confidence interval are \$7,660 and \$10,316, respectively. Comparing these amounts with the lower and upper limits of the cost-accounting-based estimate (\$13,520 and \$18,208) indicates more clearly why the null hypothesis was rejected. The correlation coefficient between experience as a route salesman and the CREPID-based values was calculated, with results similar to those found with the cost accounting method: *r* = 0.238, *p* > .10, *ns*. In other words, approximately 5.7% (0.238²) of the variability in the CREPID values of worth was explained by experience.

Examination of the 90% confidence interval of the CREPID estimate (lower limit, \$7,660; upper limit, \$10,316) indicates some degree of overlap with the 90% confidence interval derived from the global estimate (lower limit, \$9,492; upper limit, \$19,780). Table 3 presents a summary of these results.

Discussion

At first glance, the results of the data analysis seem to provide strong support for the accuracy of the Global Estimation

Model, whereas the CREPID approach appears to underestimate *SD*_y. The magnitude and direction of the relations among the three estimates of *SD*_y is obvious, even without the benefit of statistical testing. However, several limitations of the cost-accounting-based measures must be considered. The discussion that follows will be organized along the following major themes: (a) an examination of the cost-accounting-based procedure as a measure of “truth,” (b) the results associated with the Global Estimation Model, in light of prior research, (c) the results associated with the CREPID model, in light of prior research, (d) implications for accounting research, and (e) implications for future psychometric tests of the *SD*_y parameter.

Validity of the Cost-Accounting-Based Approach

Our study used a cost-accounting-based estimate of *SD*_y as a standard for comparing the behaviorally based methods for estimating *SD*_y. Therefore, conclusions about the validity (or lack of validity) of the behaviorally based approaches for estimating *SD*_y necessarily depend on the validity of the cost-accounting-based approach. Careful scrutiny of the proposed cost-accounting methodology revealed that the procedure was not based entirely on objective data. Note that the first five steps of the cost-accounting-based approach involved the use of objective, verifiable data, and involved procedures that were supportable by conventional managerial accounting theory. However, determining the degree of influence that a route salesman has over the output level on his route (Step 6) was not a completely objective process.

The approach used in the current study to determine this degree of influence could be labeled a universal, nonspecific approach. The approach was universal in that the question was phrased in terms of the “general level of influence exerted by *Route Salesmen* over sales/output on their routes.” Two major issues merit consideration with respect to Step 6 of the cost-accounting-based procedure.

First, the method for determining the influence percentages was the Delphi method. Although there is an extensive body of research supporting the procedure as an effective method for structuring a group-consensus-seeking process and for improv-

³ Interested readers may obtain a list of the 26 principal activities and the weights assigned by the judges by writing to the first author.
⁴ *t* = (*s*₁² - *s*₂²)√*N* - 2/√4*s*₁²*s*₂²(1 - *r*₁₂²).

ing the efficiency and quality of the group decision-making process (Linstone & Turoff, 1975), the Delphi method is a behavioral technique. The original objective was to estimate SD_y using a method based completely on objective, reliable, verifiable cost-accounting procedures. Integrating a behaviorally based procedure with the cost-accounting data may limit our ability to conclude that what is called the cost-accounting-based estimate of SD_y is, in fact, a measure of "truth," against which the other behaviorally based measures can be evaluated. The issue here is not the relative validity of the Delphi technique⁵ for achieving group consensus; rather, the issue centers around the propriety of incorporating a behaviorally based method into a cost-accounting-based estimate of SD_y .

Second, the percentages identified were applied to route salesmen across the board. Therefore, distinctions among levels of performance and worth were solely a function of the total output on each route. The problem is obvious: Individual differences in performance result in differing levels of influence in the home and cold-bottle markets. A more accurate approach would have involved establishing the level (or percentage) of influence exercised by each individual route salesman over the sales on his specific route. With such an approach, distinctions among levels of performance and dollar worth would be a function both of the total output on each route and of the percentage of that total attributable to the efforts of the respective route salesman.

Cravens and Woodruff (1973) outlined a stepwise regression procedure for determining criteria of sales performance that may have value in addressing this problem. They identified territory performance as a function of (a) industry market potential in the territory, (b) territory work load, (c) sales experience, (d) sales motivation and effort, (e) company experience in the territory, (f) company effort in the territory, and (g) other factors. Stepwise multiple regression was used to analyze the relation between the criterion variable (sales territory performance) and the predictor variables (industry market potential, territory work load, etc.).

Such a procedure could be integrated with the cost-accounting procedures of the current study (Steps 1–5) as a method for identifying the portion of the total contribution (as opposed to sales) attributable to a given route salesman, *if* data are available with which to measure the variables. To the extent that objective measures of predictor variables with strong relations to the criterion variable can be identified, a procedure of this nature would pose no threat to the objectivity of the remaining cost-accounting-based information used in the model. The possibilities for future studies are encouraging.⁶

We discovered an interesting fact while investigating the extreme values in the cost-accounting data set: Subsequent to the period covered by the study, the route salesman at the low end and the route salesman at the high end of the distribution were fired. The low performer was fired for nonperformance; however, the highest producing route salesman was fired for reckless driving that led to a serious accident and a lawsuit against the organization. The costs associated with this reckless behavior were not evident in the cost-accounting-based values; however, further investigation revealed that the supervisors rating this individual using CREPID took these behaviors into account. The result was low degree of agreement between the cost-account-

ing-based estimate of worth and the CREPID model estimate of worth for this individual.

Global Estimation Model

Proponents of the Global Estimation Model are sure to be encouraged by the results of the current study. Three observations seem warranted.

First, Bobko et al. (1983), in an earlier study, found that although supervisors underestimated the actual values of the three point estimates (as indicated by objective sales data), the SD_y estimate produced by the Global Estimation Model was quite accurate. The findings of the current study appear consistent with this finding. Table 3 reveals a mean worth to the organization of \$31,979 (71% of the mean worth estimate provided by the cost-accounting approach), but an estimate of SD_y of \$14,636 (92% of the estimate of SD_y provided by the cost-accounting approach).

Second, the Global estimate of SD_y was approximately 1.6 times larger than the CREPID estimate of SD_y . This finding is consistent with the outcome in a study by Weekley et al. (1985), in which the global estimate of SD_y was 1.8 times larger than it was in the CREPID approach. The consistency of the relation between the global and CREPID estimates of SD_y appears to be quite strong. Actually, there is a very plausible reason why the CREPID method should yield SD_y estimates that are considerably smaller than those produced by the global or cost-accounting methods. The CREPID approach is based on salary, not on the value of output as sold. It (CREPID) considers only the contribution of labor, not the combined contribution of labor, equipment, capital, overhead, and profit, as does a standard based upon the value of output as sold.

In the U.S. economy, wages average 57% of the value of output (Hunter & Schmidt, 1983). Hence, one would expect the CREPID SD_y estimate to average about 57% of the value of SD_y based on methods that consider the value of output as sold. As can be seen in Table 3, the CREPID SD_y estimate is exactly 57% of the cost-accounting-based SD_y , and 61% of the SD_y value produced by the Global Estimation Model. Conversely, results from both the cost-accounting procedure and Global Estimation Model should be approximately $1/0.57 = 1.75$ times as large as the CREPID SD_y value—and they are. This kind of multithreshold convergence is impressive.⁷

Three, the global estimate of SD_y was 55% of the average salary of route salesmen, somewhat larger than the 40% lower bound suggested by Hunter and Schmidt (1983), but significantly less than the 133% observed by Reilly and Smither (1985). However, examination of raters' raw estimates reveals at least two sets that were significantly greater than the other estimates, suggesting different interpretations of overall worth.

⁵ In fact, the Delphi technique has been proposed by Bobko et al. (1983) as a method for improving the accuracy of Global Estimation Model point estimates of SD_y . A subsequent study by Burke and Frederick (1984) indicated that such a procedure does, in fact, reduce within-column (percentile point) variation.

⁶ We made an attempt in the current study to develop a similar model; however, the organization was unable to provide the necessary data.

⁷ We would like to acknowledge an anonymous reviewer for pointing out these relationships among alternative methods for estimating SD_y .

This phenomenon was observed in studies by Bobko et al. (1983), Weekley et al. (1985), and Reilly and Smither (1985). However, when these outlier values are eliminated from the calculations, the estimate of the average worth of route salesmen is reduced from \$31,979 to \$26,200. This value is 99% of the actual average wage paid (\$26,585) to route salesmen. The closeness of the estimates supports the notion that most of the supervisors were using salary as the criterion for estimating worth, at least at the 50th percentile. Estimates at the 15th and 85th percentiles, however, probably were based on a standard other than salary. These are the two estimates in which supervisors assess the magnitude of individual differences among employees at the same (or similar) salaries. They do this against whatever standard they have established for the worth of the average employee. Because no such assessment of individual differences is involved in estimating the 50th percentile, it is reasonable to suspect that the 50th percentile estimates were based on salary. Anecdotal evidence makes such an explanation even more plausible. During each administration of the Global Estimation Model questionnaire, opportunity was provided for questions of clarification. In each instance, without exception, at least one rater asked whether he could use salary as a basis for making the estimates.

CREPID Model

The conservative estimate of SD_y yielded by the CREPID model is consistent with the earlier finding by Reilly and Smither (1985) that, at worst, CREPID underestimates the value of labor. Studies by Weekley et al. (1985) and Reilly & Smither found the CREPID procedure to yield a smaller estimate of SD_y than the Global Estimation Model. This finding was reinforced by the current study. Several observations with respect to the CREPID procedure merit attention.

First, the CREPID procedure produced a much tighter range of values than did the other methods. The CREPID range of values was \$36,316, as compared to cost-accounting and global ranges of \$88,320 and \$175,000, respectively (see Table 3). Excessive variability in the point estimates of worth, combined with confusion regarding interpretations of what to use as an index of individual worth, are practical problems that have been identified with use of the Global Estimation Model (Reilly & Smither, 1985). The results of the current study indicate that CREPID avoids these problems.

Second, the CREPID model produced a median value of worth (\$40,489) that was 98% of that produced by the cost-accounting-based method (\$41,166). Although median worth is not the parameter of interest in this study, the closeness of the values is worth noting.

Third, the CREPID model had the highest degree of face validity with those individuals within the organization who were familiar with all three methodologies. During the course of the study, four of the top managers and the firm's accountant became familiar with each of the procedures and the data used to provide the estimates of SD_y . These individuals were asked, individually, which of the procedures they would prefer. The unanimous choice was the CREPID procedure. Earlier, it was noted that the behaviorally based methods for estimating SD_y lack credibility among practitioners, due to the fact that they have not been validated by examining them against some mean-

ingful external criteria, such as cost-accounting outcomes. This finding seems to indicate that the credibility of the CREPID procedure may not be a significant issue.

Implications for Accounting Research

Earlier, it was noted that human resource accounting has a limited future with respect to financial reporting. Although most researchers have abandoned attempts to place a dollar valuation of human assets on the balance sheet, recent studies have focused on the decision usefulness of human resource information. Shimerda and Pufahl (1983) identified five studies in which the provision of human resource data resulted in decisions that were significantly different from those that would have been made on the basis of conventional financial data alone. A study by Hendricks (1976) found that stock investment decisions were affected by including human resource accounting information with conventional financial statements. The current study presents another challenge for accountants: that of developing an objective, verifiable, and reliable method for estimating the dollar value of job performance levels. As noted earlier, the cost-accounting systems of organizations have not established job performance level as a cost object, inasmuch as estimation of the dollar differences in levels of job performance traditionally has not been a cost objective. Given the decision usefulness of human resource information, and given the specific informational need described in this article (i.e., the ability to produce a valid, cost-accounting-based estimate of SD_y), the implication of the current study is clear: Researchers within the accounting profession must develop an objective, verifiable, and reliable method for estimating the standard deviation of job performance in dollars (SD_y). Although behaviorally based methods have been developed for the estimation of SD_y , Weekley et al. (1985) noted that without an appropriate outside criterion against which to validate the behaviorally based methods for estimating SD_y , one is unable to support strongly one method over the others. The accounting sector must provide that outside criterion. The current study merely provides a start in this direction.

Implications for Future Tests of the SD_y Parameter

Future studies evaluating the SD_y parameter should attempt to eliminate the sources of within-column variability of the Global Estimation Model. Future studies should focus on the identification and elimination of the sources of rater confusion in the completion of the Global Estimation Model questionnaire. There is a need for researchers to examine conditions under which the CREPID model underestimates (or overestimates) SD_y . There is a need for more studies that compare and contrast behaviorally based estimates of SD_y with external, objective measures of SD_y . Finally, future studies should continue a call to the accounting profession to develop objective, verifiable, and reliable methods for estimating the SD_y parameter.

Despite the fact that we have focused exclusively in this study on the SD_y parameter, other parameters in the general utility equation also deserve attention, for they too are estimates. Much work needs to be done in developing refined estimates of the size of the samples selected or trained, the mean time selectees stay on the job, or the mean duration of a training effect,

and the selection ratio. That kind of research can lead only to better decisions about personnel programs, and better decision making is what utility analysis is all about.

References

- Baker, B. O., Hardyck, C. D., & Petrinoich, L. F. (1966). Weak measurements versus strong statistics: An empirical critique of S. S. Stevens' proscriptions on statistics. *Educational and Psychological Measurement*, 26, 291-309.
- Bedford, N. M., Brummet, R. L., Churchill, N. C., Fertig, P. E., Morrison, R. H., Salmonson, R. F., Sorter, G. H., Vance, L. L., & Zlatkovich, C. T. (1966). *A statement of basic accounting theory*. Sarasota, FL: American Accounting Association.
- Blum, M. L., & Naylor, J. C. (1968). *Industrial psychology: Its theoretical and social foundation* (Rev. ed.). New York: Harper & Row.
- Boatsman, J. R., Demski, J., Kennelly, J. W., Larson, K. D., Revsine, L., Staubus, G. J., Sterling, R. R., Weygandt, J. J., & Zeff, S. (1977). *Statement on accounting theory and theory acceptance*. Sarasota, FL: American Accounting Association.
- Bobko, P., Karren, R., & Parkington, J. J. (1983). Estimation of standard deviations in utility analysis: An empirical test. *Journal of Applied Psychology*, 68, 170-176.
- Boneau, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin*, 57, 49-64.
- Boudreau, J. W. (1983). Economic considerations in estimating the utility of human resource productivity improvement programs. *Personnel Psychology*, 36, 551-576.
- Brogden, H. E. (1949). When testing pays off. *Personnel Psychology*, 2, 171-185.
- Burke, M. J., & Frederick, J. T. (1984). Two modified procedures for estimating standard deviations in utility analysis. *Journal of Applied Psychology*, 69, 482-489.
- Cascio, W. F. (1980). Responding to the demand for accountability: A critical analysis of three utility models. *Organizational Behavior and Human Performance*, 25, 32-45.
- Cascio, W. F. (1987a). *Applied psychology in personnel management* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Cascio, W. F. (1987b). *Costing human resources: The financial impact of behavior in organizations* (2nd ed.). Boston: Kent.
- Cascio, W. F., & Ramos, R. A. (1986). Development and application of a new method for assessing job performance in behavioral/economic terms. *Journal of Applied Psychology*, 71, 20-28.
- Cherrington, J. O., Hubbard, E. D., & Luthy, D. H. (1985). *Cost and managerial accounting*. Dubuque, IA: Wm. C. Brown.
- Cravens, D. W., & Woodruff, R. B. (1973). An approach for determining criteria of sales performance. *Journal of Applied Psychology*, 57, 242-247.
- Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.
- Dalkey, N. C., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management Science*, 9, 458-467.
- DeAngelo, L. E. (1982). Unrecorded human assets and the 'hold up' problem. *Journal of Accounting Research*, 20, 272-274.
- Dittman, D. A., Juris, H. A., & Revsine, L. (1976). On the existence of unrecorded human assets: An economic perspective. *Journal of Accounting Research*, 14, 49-65.
- Dittman, D. A., Juris, H. A., & Revsine, L. (1980). Unrecorded human assets: A survey of accounting firms' training programs. *The Accounting Review*, 55, 640-648.
- Hardyck, C. D., & Petrinoich, L. F. (1969). *Statistics for the behavioral sciences*. Philadelphia, PA: Saunders.
- Hendricks, J. A. (1976). The impact of human resource accounting information on stock investment decisions: An empirical study. *The Accounting Review*, 51, 292-304.
- Horngren, C. T. (1982). *Cost accounting: a managerial emphasis* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Hunter, J. E., & Schmidt, F. L. (1983). Quantifying the effects of psychological interventions on employee job performance and work-force productivity. *American Psychologist*, 38, 473-478.
- Landy, F. J., Farr, J. L., & Jacobs, R. R. (1982). Utility concepts in performance measurement. *Organizational Behavior and Human Performance*, 30, 15-40.
- Linstone, H. A., & Turoff, M. (1975). *The Delphi method: Techniques and applications*. Reading, MA: Addison-Wesley.
- McNemar, Q. (1969). *Psychological statistics* (4th ed.). New York: Wiley.
- Naylor, J. C., & Shine, L. C. (1965). A table for determining the increase in mean criterion score obtained by using a selection device. *Journal of Industrial Psychology*, 3, 33-42.
- Reilly, R. R., & Smither, J. W. (1985). An examination of two alternative techniques to estimate the standard deviation of job performance in dollars. *Journal of Applied Psychology*, 70, 651-661.
- Roche, W. J., Jr. (1961). The Cronbach-Gleser utility function in fixed treatment employee selection. *Dissertation Abstracts International*, 22, 4413. (University Microfilm No. 62-1570)
- Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). Impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology*, 64, 609-626.
- Schmidt, F. L., Mack, M. J., & Hunter, J. E. (1984). Selection utility in the occupation of U.S. Park Ranger for three modes of test use. *Journal of Applied Psychology*, 69, 490-497.
- Shimerda, T. A., & Pufahl, D. R. (1983). The effects of human resource accounting data on decision making. *Cost and Management*, 4, 41-45.
- Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection. *Journal of Applied Psychology*, 39, 565-578.
- Weekley, J. A., Frank, B., O'Connor, E. J., & Peters, L. H. (1985). A comparison of three methods of estimating the standard deviation of performance in dollars. *Journal of Applied Psychology*, 70, 122-126.

Received June 30, 1986

Revision received February 24, 1987

Accepted April 14, 1987 ■

A Model of Hiring Decisions in Real Employment Interviews

Susan M. Raza and Bruce N. Carpenter
University of Tulsa

Past research on variables affecting hiring decisions has emphasized the role of applicant and interviewer demographics. However, recent studies have questioned the generalizability of findings from laboratory interviews to real interviews. In this article, a model of demographics and interviewing decisions is proposed and tested with actual employment interviews. Industrial interviewers ($N = 8$) provided demographic data concerning themselves and applicants ($N = 171$), rated applicants on widely studied attributes, and made two hiring decisions. The data support the model that interview outcomes are directly dependent on the more logically relevant variables, such as skill. Furthermore, the influence of demographics is modest and less important than other variables.

Interviews continue to be the most widely used selection device, although studies have generally found interviews to have low reliability and validity and to be subject to various biases (Arvey, 1979; Arvey & Campion, 1982). Unfortunately, recent research concerning the paradigms on which these findings were based suggest that earlier findings may not be generalizable to actual employment settings (e.g., Gorman, Clover, & Doherty, 1978). Therefore, in this study we reexamine the relative influence of demographic information and interviewer rating of applicant characteristics on employment interview outcomes. We do not address the issue of the reliability and validity of interviews—an important, though separate issue. By simultaneously examining the interviewer rating of applicant characteristics and the demographic variables, we address the question of their relative as well as absolute influence upon outcomes. We propose and test a structural causal model explaining how potentially important interviewer and interviewee characteristics contribute to outcomes.

Design Concerns

Most investigators have studied interviewing in artificial situations. They have used (a) interviewers who were trained for the purpose of conducting research rather than using professional interviewers and (b) artificial jobs rather than actual job openings. Furthermore, they have frequently used the "paper-people paradigm," presenting the applicant through transcripts of interviews, application forms, resumes, test results, or photographs. To assume that such results generalize to the actual interview situation is highly questionable.

Several studies suggest that the paper-people paradigm yields results that differ from those using actual interviews. For example, Gorman et al. (1978) compared the two methods by asking graduate students and experienced interviewers to make judgments based on interviews or objective data. They found that interviewers' judgments substantially differed as a function of

the stimulus presented and concluded that we will learn little about actual interviewing from studies using the paper-people paradigm. This suggests limited generalizability from artificial to actual interviewing situations.

A study of graduate recruitment interviews (Keenan, 1977) yielded a strong relationship between liking the candidate and the overall evaluations, which held up when the effect of ratings of the candidate's intelligence were held constant. The study suggests that there is an affective component to actual interviews that is not operable in paper-people paradigm research.

Okanes and Tschirgi (1978) compared preinterview judgments based on paper data to postinterview judgments. Using the categories of *unable to determine*, *probably recommend*, and *probably not recommend*, they found significant shifts from the *unable to determine* and *probably recommend* categories. Most shifts were to the *probably not recommend* category, indicating that the interviewers were more negative following the actual interviews. The study does not address which variables are associated with the shifts, but does indicate that judgments based on paper-people differ from those based on actual interviews.

Finally, Dipboye, Fontenelle, and Garner (1984) used three groups of college student interviewers to study the effects of using (a) an application and an interview, (b) an application alone, and (c) an interview alone. They found significant differences in the accuracy and reliability of evaluations as a function of the stimulus material; interrater reliability was highest for the interview-only condition and lowest for the application-only condition. Additionally, interviewers in the interview-only condition more accurately estimated the applicants' self-described personalities than did interviewers in the other two conditions. This study indicates that results differ as a function of the research paradigm.

Influence of Demographics

Prior research suggests that interview outcomes are associated with various interviewer and applicant characteristics. For example, younger candidates received higher evaluations than older candidates (Avolio, 1982; Haefner, 1977); female interviewers rated all applicants higher than did male interviewers

Correspondence concerning this article should be addressed to Bruce N. Carpenter, Department of Psychology, University of Tulsa, Tulsa, Oklahoma 74104.

(London & Poplawski, 1976); women generally received lower evaluations than men (Dipboye, Fromkin, & Wiback, 1975); interviewers rated more positively those applicants whom they liked and whom they considered more intelligent (Keenan, 1977); and qualified and attractive candidates were more likely to be recommended for hiring regardless of sex (Dipboye et al., 1975).

Current Study

The evidence has consistently shown that interview decisions appear to be largely determined by biases regarding irrelevant demographic characteristics. However, such findings may reflect the artificial nature of studies that draw such conclusions and that have the tendency to examine bias variables alone rather than with variables reflecting more relevant characteristics of the applicant. For example, if interviewers are primarily given data on irrelevant demographics, it may be natural for them to rely on biases when asked to make employment decisions because that is what the situation allows. We therefore conducted a study that attended to these concerns.

We hypothesized that demographic variables will have limited influence on interview outcomes in interviews conducted by professional interviewers, because opportunity exists to evaluate more relevant characteristics and such interviewers have more training, experience, and motivation to control inappropriate influences. Therefore, we examined the operation of applicant and interviewer sex and age on interviewer decision variables. We included an additional nonrating variable, *type of job*, because it was thought to be potentially related to outcome through the job status or through sex or age stereotyping associated with the type of job. For example, people applying for high-level jobs may receive higher hiring recommendations as a result of the job's status.

We proposed that interviewers evaluate applicants on dimensions other than demographics and reach outcome decisions based on these other variables, most of which have also been studied earlier but not clearly related in the decision process. Therefore, we also examined the relationship of decision variables to interviewer ratings of physical attractiveness, personal liking, intelligence, and skill level for the job. We considered only the interviewer's ratings of these variables, rather than more objective measures of applicants, such as IQ, because this study examines the effects and interrelationships of variables on interview outcomes rather than interview validity. Thus, ratings are viewed as reflections of the interviewer's decision making, even when objective measures are available.

Decision Variables

Two decision variables were used in this study. The interviewer first rated the applicant on degree of recommendation for hiring, a variable commonly used in interviewing studies. The interviewer also rated what kind of employee the applicant would make, a variable we could not find used elsewhere in the literature. It is likely that an applicant may have the characteristics that would yield a good employee, but not have the skills and abilities appropriate to the available job. This second variable allowed interviewers to rate applicants' general employ-

ability, independent of how well the applicant matched the particular opening that was under consideration.

Hypothesized Model

We hypothesized a priori how the various rating variables would be related to one another so as to yield the interview decisions. The most important influences or those that we expected to be inadequately represented by intervening variables were modeled as having direct relationships with the decision variables. Either influences of lesser importance, or influence that we expected to be primarily subsumed by intervening variables, or both sets of influences, were modeled as having indirect relationships. Under our model, the hirability of applicants (hiring decision) is directly influenced by judgments of the applicants' employability and skill. That is, we expected that those recommended for hiring would be judged as potentially good general employees and would have the requisite skills for the particular position. We expected that the influences of the rating of personal liking and intelligence, and that portion of the skill rating that may result from judging that the applicant is able to readily learn specific skills, would be adequately represented by the general employability decision. Thus, personal liking, intelligence, and skill are modeled as indirectly affecting the hiring decision through the employability decision. The general employability of an applicant is directly influenced by judgments of personal liking (likability), intelligence, and skill. That is, people who are considered bright, capable, and pleasant are those most expected to be judged as making good employees. Physical attractiveness ratings were expected to be adequately represented by likability ratings. Thus, we hypothesized that physical attractiveness ratings indirectly influence the employability decision through likability.

Similarly, we hypothesized that rating variables would have direct and indirect effects on one another. Skill ratings are directly influenced by judgments of the applicant's intelligence. That is, people perceived as brighter are expected to be rated as more skillful. Intelligence ratings are directly influenced by likability. And finally, likability ratings are directly influenced by judgments of attractiveness.

Our understanding of actual employment interviews leads us to conclude that demographic and job type variables have only a small influence on interview outcomes. Therefore, we propose that they have only indirect relationships to those outcomes through the rating variables. We initially hypothesized that demographic and job type variables will indirectly influence the decision process at least through their direct relationship to physical attractiveness ratings. Because we have no theoretical basis for predicting any other possible direct relationships with rating variables, we decided to retain only those additional paths that were significant at the .05 level. This complete path model of demographic, rating, and decision variables is diagrammed in Figure 1.

Method

Subjects

The 171 interviews used in this study were conducted during the summer of 1985 in a major city in the southwestern United States. However,

the variety of interviewer and applicant characteristics represented by the sample affords considerable generalizability. The unit of analysis is the interview; the sample size has a power of .9 to detect a medium ($r = .30$) effect at the .01 level of significance for a two-tailed test.

Interviewer characteristics. Eight industrial interviewers conducted 171 initial job interviews during the course of their normal duties. The interviews were not standardized or structured other than by the particular interviewer. The application blank with demographic information and job applied for was available to the interviewer prior to the interview. Interviewer ages ranged from 25 to 45 with a median of 31. Four of the interviewers were women and four were men.

The interviewers represented a variety of industries, and the percentage of interviews conducted by industry type was fairly evenly distributed according to the number of interviewers. One each worked for a major hospital (12% of interviews), a major energy corporation (13%), an executive search firm (14%), and a general employee search firm (12%). The remaining four interviewers worked for the city (population of about 500,000) (49%).

Applicant characteristics. The 171 applicants were nearly equally divided as to sex (53% men, 47% women) and represented a wide age range (from age 18 to 62, with a median of 31). The distribution of applicant age was slightly skewed toward the lower end. The applicants applied for a wide variety of jobs (e.g., managerial, clerical, and semi-skilled labor), and the percentage of applicants per job was fairly evenly distributed.

Procedure

Immediately following each interview the interviewers filled out a one page questionnaire. The questionnaire requested (a) *demographic information*: interviewer age and sex, applicant age and sex, job for which applicant is applying; (b) *ratings of the applicant*: intelligence, physical attractiveness, likability (how much interviewer likes applicant), and skill level for the job; and (c) *ratings on two decision variables*: hirability (degree of hiring recommendation) and employability (the general acceptability of the applicant as an employee). All ratings were single items on 5-point scales. Therefore, reliability of ratings cannot be assessed. Thus, the reader is cautioned that differences in magnitude of relationships may be, in part, due to differences in reliabilities. A list of the variables according to category is presented in Table 1. The jobs for which the applicants applied were categorized and ordered on four separate dimensions. In accordance with the *Dictionary of Occupational Titles* (DOT; U.S. Department of Labor, 1977), each job was ordered according to the level of data, people, and things. (Level of things refers to job requirements for use and physical manipulation of objects.) To simplify the presentation of results we reversed the DOT's coding

system so that high numbers reflect high levels of data, people, and things. Employer-provided information was used to order the jobs according to required education and experience. This was done because interviewers may be more generally aware of the educational and experiential requirements of a job than of the level of data, people, and things associated with a job. They may thus see a job as having high status as a result of its requiring a high level of education and experience rather than as a result of its being associated with high levels of handling data, people, and things.

Factor analysis of the four categorizations yielded two factors. The first factor, consisting of jobs as ordered by data, people, and education and experience, accounted for 68% of the total variance. The second factor, consisting of jobs as ordered by things, accounted for 29% of the total variance. An examination of the correlation matrix revealed that jobs as ordered by things is inconsistently correlated with the remaining three job categories and has nonsignificant correlations with all rating variables; it was therefore dropped from the study. Instead of using the individual variables of jobs ordered as to data, people, and education and experience, we used scores based on unit weighting of the three variables with large loadings ($>.90$) on Factor 1 for the remainder of the analyses.

Results

Interview Variables

Age of interviewer and applicant. There were no significant relationships between the age of the interviewer and the applicant rating or decision variables. However, the older applicants were rated significantly lower in intelligence by the entire sample, $r(171) = -.30, p < .001$, and by male interviewers, $r(71) = -.57, p < .001$. Also, older applicants were rated lower in attractiveness by female interviewers, $r(83) = -.33, p < .01$. For the entire sample the older applicants received lower hiring recommendations, $r(171) = -.20, p < .01$. However, when the sample was divided according to the sex of the interviewer, the negative correlations between applicant age and hiring recommendations held only for men (female interviewers, $r(83) = -.12; p = .31$; male interviewers, $r(71) = -.25, p < .01$).

Sex of interviewer and applicant. The female interviewers rated all applicants significantly higher than did male interviewers on intelligence, $t(159) = 4.60, p < .001$, attractiveness, $t(160) = 5.89, p < .001$, likability, $t(160) = 2.89, p < .01$, and skill, $t(160) = 4.38, p < .001$. Hiring decisions were also higher

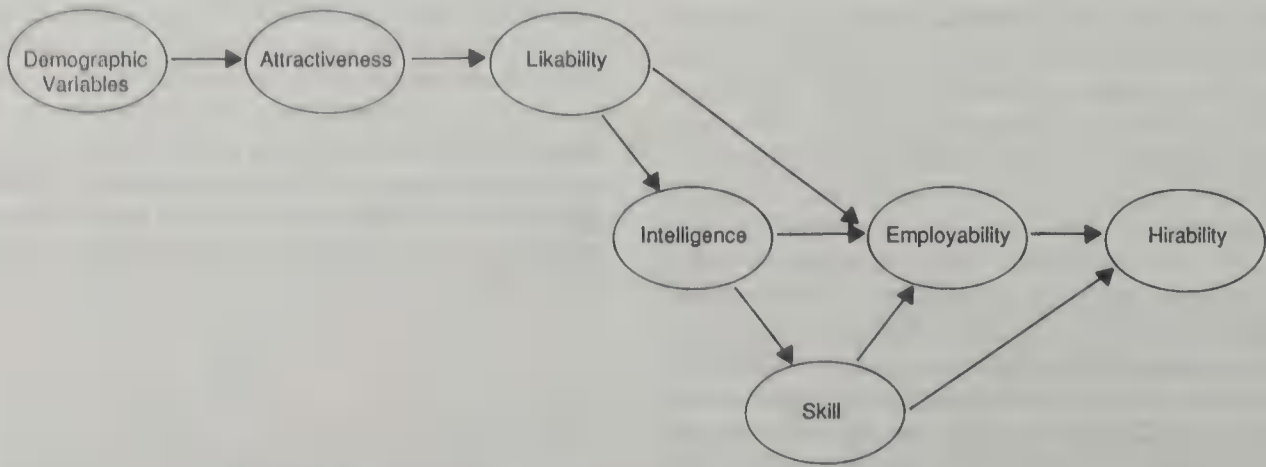


Figure 1. Proposed model of specific rating variables and outcome variables.

Table 1
Study Variables Organized by Variable Type

Variable	Variable
Demographics	Interviewer ratings of applicant
Interviewer	Attractiveness
Age	Likability
Sex	Intelligence
Applicant	Skill
Age	Interviewer decisions regarding applicant
Sex	Hirability
Job type (factor score of data people, and education/ experience)	Employability

for female versus male interviewers, $t(159) = 3.60, p < .001$. Male interviewers gave female applicants higher attractiveness ratings than they gave male applicants, $t(69) = 2.54, p < .01$, and female interviewers rated male applicants significantly higher on attractiveness ratings than did male interviewers, $t(81) = 2.82, p < .01$. There was no significant difference in employability ratings.

The female applicants received significantly higher ratings than males on intelligence, $t(160) = 3.32, p < .01$, attractiveness, $t(160) = 4.86, p < .001$, and skill, $t(160) = 2.85, p < .01$. All other comparisons between applicant sex and rating and decision variables were nonsignificant.

Job type. People who applied for jobs that are rated high in data, people, and education and experience requirements received higher ratings of intelligence, $r(171) = .30, p < .001$, attractiveness, $r(171) = .32, p < .001$, and skill, $r(171) = .28, p < .001$. They also received higher employability recommendations, $r(171) = .19, p < .01$. Job type was not significantly associated with degree of hiring recommendation.

Decision variables. Employability and hirability were positively correlated, $r(171) = .65, p < .001$, but differential patterns of correlations suggest they are not so redundant as to be seen as a single interview outcome. Not surprisingly, hirability and employability decisions were positively correlated with intelligence, attractiveness, likability, and skill ratings (see Table 2). Employability was most strongly associated with intelligence and likability, whereas hirability was most strongly associated with skill. The magnitude of the associations did not significantly differ for male and female interviewers.

Table 2
Correlations Between Applicant Ratings and Decision Variables for the Entire Sample and for Male and Female Interviewers

Ratings	Employability				Hirability			
	Total	Male	Female	Z	Total	Male	Female	Z
Intelligence	.66**	.66**	.65**	.22	.45**	.44**	.38**	.43
Attractiveness	.36**	.34*	.32*	.07	.30**	.10	.30**	1.31
Likability	.62**	.54**	.65**	1.06	.48**	.36*	.48**	.94
Skill	.53**	.45**	.57**	1.00	.70**	.77**	.68**	1.13

Note. Zs represent Z test for comparing correlations of male versus female samples.
* $p < .01$. ** $p < .001$.

Path Analysis

Analyses revealed that each of the paths proposed in the a priori model was significant. Three additional paths between job type and intelligence ratings, applicant age and intelligence ratings, and interviewer sex and skill ratings, were also significant at the .05 level and were added. The overidentified model, with the path coefficients, R^2 s, and e 's, is presented in Figure 2. The five exogenous variables are intercorrelated, but the conventional curved, double-headed arrows are omitted from the figure solely for figural clarity.

As suggested by Specht (1975) we tested the overall goodness of fit of both the overidentified and just-identified models using the appropriate generalized multiple correlation coefficients and W tests of significance. For the overidentified model $M(171) = .96, W(16) = 491.82, p < .001$, and for the just-identified model $R^2_m = .97, W(45) = 432.16, p < .001$. These results indicate that both models fit the data very well.

We then tested our 16 path overidentified model against the 45 path just-identified model using the Q goodness-of-fit statistic and the W test of significance suggested by Pedhazur (1982), $Q(171) = .77, W(29) = 36.46, .10 < p < .25$. These results suggest that our more parsimonious model provides a fit of the data that is not significantly poorer than the just-identified model and that the null hypothesis of no difference between the two models cannot be rejected.

To further test our model we reproduced the original correlation matrix using the formulae suggested by Pedhazur (1982). Table 3 presents the original correlations in the lower triangle and the reproduced correlations in the upper triangle. Z tests of the significance of difference between the original and reproduced correlations revealed no significant differences. The greatest difference between the original and reproduced correlations was for the correlations between likability and hirability, $Z = 1.34$ (mean $Z = .42$).

Impact of Demographic on Nondemographic Variables

To investigate the impact of the demographic variables on the nondemographic variables we conducted a series of regression analyses predicting each of the nondemographic variables first from the demographic variables alone, second from the nondemographic variables postulated by our overidentified model, and finally from both sets of variables together. Additionally, we

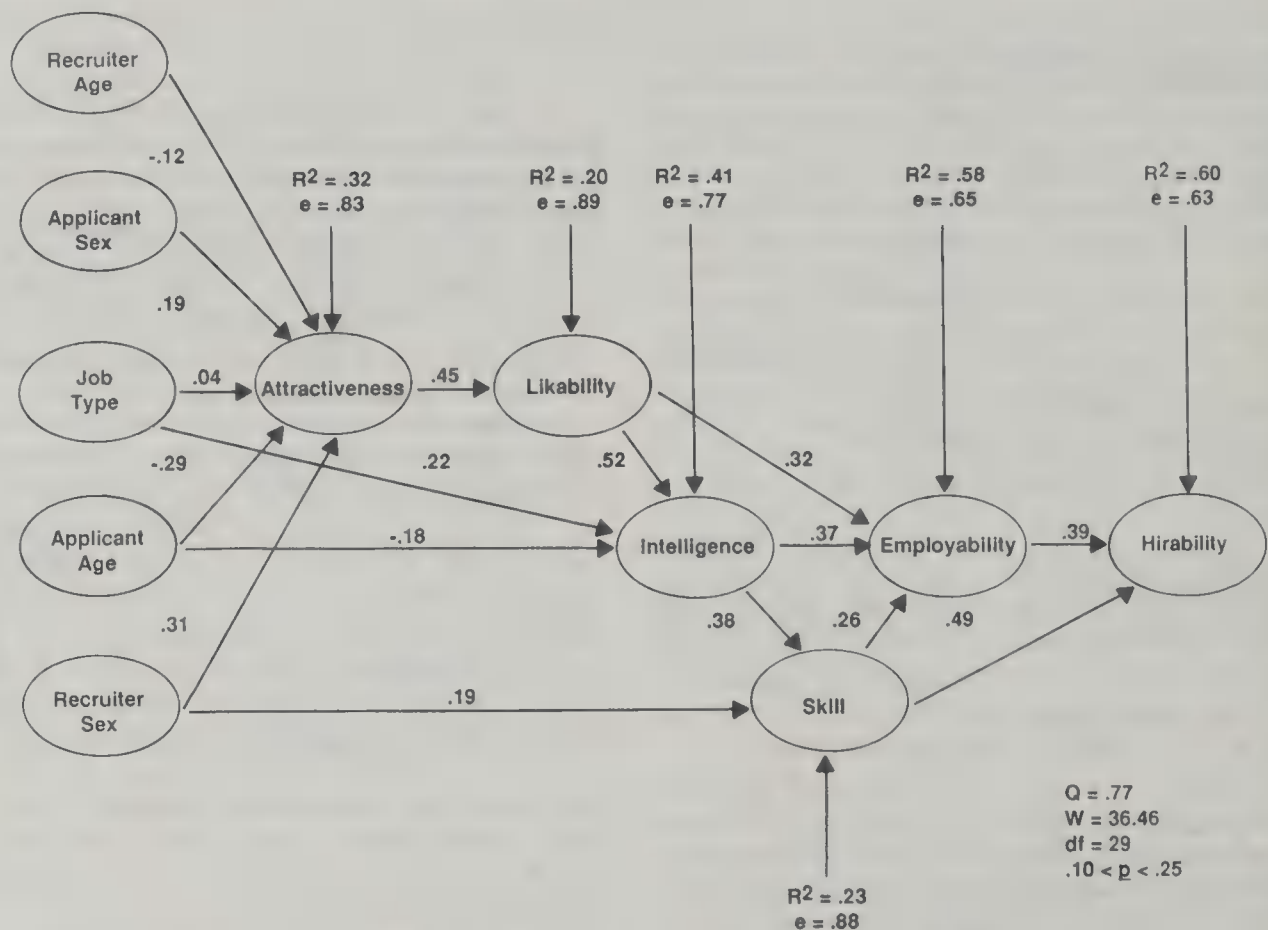


Figure 2. Overidentified model of specific rating variables and outcome variables. (The conventional curved, double-headed arrows showing that the exogenous variables are intercorrelated are omitted solely for clarity of presentation.)

computed the effect coefficients of all relevant variables for each of the endogenous variables and compared their relative size for the outcome variables.

Attractiveness. Predicting attractiveness from the demographic variables alone yielded $R^2(171) = .32$, $p < .001$, indicating that demographic variables have significant utility for predicting attractiveness.

Likability. Entering only the demographic variables yielded an $R^2(171) = .06$, $p > .05$, indicating that the demographic variables can explain a maximum of 6% of the variance in likability ratings. Entering only attractiveness, as required by the model, yielded an $R^2(171) = .20$, $p < .001$, and entering both sets of variables together yielded an $R^2(171) = .21$. The attractiveness rating, therefore explained 15% of the variance in the likability ratings beyond that explained by demographics and is relatively more important than the demographic variables for predicting likability.

Intelligence. Entering only the demographic variables yielded an $R^2(171) = .20$, $p < .001$, indicating a maximum of 20% of predictive power for intelligence ratings. Entering only likability yielded an $R^2(171) = .31$, and entering both sets of variables together yielded $R^2(171) = .42$. Likability can then explain at least 22% of the variance in intelligence ratings beyond that of the demographics and may be only slightly more important than the demographic variables for predicting intelligence. This is reflected in our overidentified model, which has

two significant direct paths from the demographic variables to the intelligence ratings (job type and applicant age).

Skill. The demographic variables alone yielded an $R^2(171) = .14$, $p < .001$, for a maximum of 14% predictive power. Entering only the intelligence ratings yielded an $R^2(171) = .19$, and entering both sets of variables together yielded an $R^2(171) = .24$. Intelligence ratings can explain at least 10% of the variance in the skill ratings beyond that of the demographics and may not, therefore, be more important than the demographic variables in predicting skill. Our model indicates that at least the gender of the interviewer is important for predicting the skill ratings.

Employability. Entering only the demographic variables yielded an $R^2(171) = .05$, ns , so that demographic variables have a maximum of 5% predictive power. Entering likability, intelligence, and skill as set by the model yielded an $R^2(171) = .58$, and entering both sets together yielded an $R^2(171) = .60$. Therefore, likability, intelligence, and skill together can, controlling for the effect of demographics, explain at least 55% of the variance in employability outcome and are relatively more important for predicting employability than the demographic variables.

Hirability. The demographic variables alone yielded an $R^2(171) = .10$, $p < .01$, for a maximum of 10% predictive power. Skill and employability yielded an $R^2(171) = .60$, $p < .001$, and both sets of variables together yielded an $R^2(171) = .61$. Skill

Table 3
Original and Reproduced Inter correlations of Specific Rating Variables and Outcome Variables

Variable	1	2	3	4	5	6	7	8	9	10	11
1. Interviewer age	—					.04	.02	.01	.09	.03	.05
2. Applicant sex	.26	—				.36	.16	.19	.17	.17	.15
3. Job type	-.03	.44	—			.32	.15	.34	.20	.23	.19
4. Applicant age	.07	-.07	-.25	—		-.37	-.17	-.32	-.15	-.21	-.16
5. Interviewer sex	.43	.50	.38	-.15	—	.42	.19	.21	.27	.21	.21
6. Attractiveness	.04	.36	.32	-.37	.42	—	.45	.37	.21	.34	.24
7. Likability	.09	.14	.07	-.12	.22	.45	—	.58	.25	.60	.36
8. Intelligence	.06	.24	.30	-.30	.34	.44	.55	—	.42	.67	.46
9. Skill	.10	.22	.28	-.17	.32	.21	.34	.44	—	.50	.69
10. Employability	.05	.15	.19	-.13	.16	.36	.62	.66	.53	—	.63
11. Hirability	.13	.13	.14	-.20	.27	.30	.48	.44	.70	.65	—

Note. Original correlations are in the bottom triangle, reproduced correlations are in the top triangle.

and employability alone can explain at least 51% of the variance in hirability and are, therefore, relatively more important than the demographic variables for predicting hirability.

Effect coefficients. Our more parsimonious model, with demographic variables only indirectly affecting interview outcomes, fits the data as well as the just-identified model. This fact, together with the results of the foregoing regression analyses, strongly suggests that demographic variables have a limited effect on interview outcomes. However, demographic variables do significantly affect several of the applicant rating variables and, therefore, indirectly affect the interview outcomes. To explore the degree of effect, we computed effect coefficients as suggested by Lewis-Beck and Mohr (1976). These are presented in Table 4 for both rating and outcome variables, although our discussion will be limited to those for the outcome variables.

The effect coefficients are separately examined for each outcome variable, as they may not be compared across outcomes. The relative effects of the coefficient are given by their absolute size. For the employability outcome the interviewer's personal liking of the applicant carries the greatest weight, followed closely by intelligence ratings. Ratings of skill and physical attractiveness are of lesser importance, and the demographic variables are least important. Similarly, skill ratings, followed by employability decisions, intelligence ratings, and ratings of

physical attractiveness, are more important for the hirability outcome than are demographic variables.

Discussion

Comparison of Results With Previous Studies

The results, consistent with paper-people paradigm results, suggest that applicant ratings, and occasionally hiring decisions, are associated with demographic variables and that applicant ratings are positively associated with interview outcomes. However, only some of the conclusions of previous studies are supported by this investigation of actual interviews.

Applicant age. The few studies that investigated the relationship of applicant age and interview outcomes (Avolio, 1982; Haefner, 1977) found that older applicants receive lower overall evaluations. In the present study older applicants were not uniformly rated lower than younger applicants, but they were differentially viewed by male and female interviewers. Older and younger applicants were rated similarly for likability, skill, and employability. Male interviewers gave older applicants lower intelligence ratings and lower hiring recommendations, whereas female interviewers gave older applicants lower attractiveness ratings. Because no data were collected as to job re-

Table 4
Effect Coefficients for Endogenous Variables

Predictor variables	Predicted variables					
	Attractiveness	Likability	Intelligence	Skill	Employability	Hirability
Interviewer age	-.12	-.05	-.03	-.01	-.03	-.02
Applicant sex	.19	.09	.05	.02	.05	.03
Job type	.04	.02	.24	.09	.12	.09
Applicant age	-.30	-.13	-.25	-.09	-.15	-.11
Interviewer sex	.31	.14	.07	.22	.13	.16
Attractiveness	—	.45	.23	.09	.25	.14
Likability		—	.52	.20	.56	.32
Intelligence			—	.38	.47	.37
Skill				—	.26	.59
Employability					—	.39

quirements, our results do not address the question of unfair discrimination, but do suggest the existence of differential rating as a function of the interviewer's sex.

Interviewer and applicant. Previous research suggests that female interviewers give higher applicant ratings than male interviewers (London & Poplawski, 1976) and that female applicants receive lower ratings than male applicants (e.g., Dipboye et al., 1975). The present study concurs that female interviewers give higher specific ratings and higher hiring recommendations than males; however, female and male interviewers do not significantly differ in employability ratings nor in the magnitude of associations between their specific ratings and the decision variables. In contrast to previous studies, female applicants in this study receive higher ratings than male applicants for intelligence, attractiveness, and skill and do not significantly differ from males in likability and outcome variables.

Liking, intelligence, skill, and attractiveness. The results of the studies by Keenan (1977) and Dipboye et al. (1975) suggest that an applicant receives higher hiring recommendations and general evaluations if he or she is liked by the interviewer and is seen as intelligent, skillful, and attractive. The results of the present study agree that higher specific ratings are associated with higher decision variables, but strongly suggest that the magnitude of the associations differs. For example, skill rating is more strongly associated with degree of hiring recommendation than are attractiveness, intelligence, or likability ratings.

Job Type and Employability

This study investigated the effects of two variables not previously studied: job type and employability. We thought it possible that hiring recommendations may be related to job type, perhaps through the status generally accorded particular positions. We learned that although people applying for jobs high in data, people, and education and experience requirements receive higher intelligence, attractiveness, likability, skill, and employability ratings, they do not receive higher hiring recommendations. Perhaps the simplest explanation is that people applying for such positions are better qualified for employment generally, but the expectations for hiring someone to fill such positions are also greater. Thus, job type may serve as a moderator variable.

We expected that an applicant may have qualities that would allow him or her to be seen as a potentially good employee, but not have the necessary skills for the available job that would yield a high hiring recommendation. Employability and hirability are correlated with one another, but the path analysis and the differential magnitude of associations between the specific rating variables and the employability and hirability variables support our model's separate treatment of them. We also learned that male and female interviewers do not significantly differ in the magnitude of associations of specific rating variables with the employability and hirability variables. We therefore conclude that employability and hirability are distinct interview outcomes and that the sex of the interviewer does not influence how specific rating variables contribute to interview outcomes.

Model of Demographic, Rating and Outcome Variables

We proposed that demographic variables directly affect attractiveness ratings, that attractiveness ratings directly affect likability ratings, that likability ratings directly affect intelligence ratings, that intelligence ratings directly affect skill ratings, that likability, intelligence, and skill ratings directly affect employability, and that skill and employability directly affect hirability. Results generally support our a priori model, although they reveal significant direct effects of job type and applicant age on intelligence ratings and of interviewer sex on skill ratings. With the addition of those three significant direct effects, the overidentified model parsimoniously and adequately explains the data of this investigation. Although the fit of our model to the data is moderately high, our model is based solely on our theoretical considerations of how the various variables work together and on the data of this study. Further studies applying the model to new data using actual interviews are necessary in order to bolster confidence in the model.

Additionally, we proposed that demographic variables would add little to our understanding of the outcome variables, but may contribute somewhat to our understanding of the specific rating variables endogenous to the model (attractiveness, likability, intelligence, and skill). Thus, we proposed that demographic variables are less important in actual interview outcomes than had been suggested by previous studies. The results of this study are very supportive of our proposal and indicate that the demographic variables of interviewer and applicant age and sex and job type are far less important to our understanding of interview outcomes, as measured by employability and hirability, than is commonly believed. Demographic variables do contribute to our understanding of attractiveness, likability, intelligence, and skill ratings. However, with the exception of that for attractiveness, their explanatory power is modest and their effect coefficients are smaller than those for the relevant non-demographic variables. Again, as our conclusions are based on a single study, replication with other samples of actual interviews is extremely important.

In summary, this study suggests that there are some important differences between results obtained from paper paradigm research and research using actual interviews. The role of demographic variables for predicting interview outcomes and interview ratings appears more modest and limited than is commonly believed. Furthermore, the applicant's intelligence and likability (social skills) appear to be most important for being rated as a potentially superior employee, but the skill rating becomes important when determining the hiring recommendation. Continued use of paper-paradigm designs to study employment interviews is probably inappropriate.

Industries are likely to continue to use interviews to make selection decisions, and it is possible that interviews are less biased than previously thought. It remains for additional research investigating actual interviews across a wide variety of jobs and interviewing situations to support this conclusion further and to clarify the validity of interview ratings.

References

- Arvey, R. D. (1979). Unfair discrimination in the employment interview: Legal and psychological aspects. *Psychological Bulletin*, 86, 736-763.

- Arvey, R. D., & Campion, J. E. (1982). The employment interview: A summary and review of recent research. *Personnel Psychology*, 35, 281-322.
- Avolio, B. J. (1982). Age stereotypes in interview evaluation contexts. *Dissertation Abstracts International*, 42, 3020B. (University Microfilms No. 81-29, 504)
- Dipboye, R. L., Fontenelle, G. A., & Garner, K. (1984). Effects of pre-viewing the application on interview process and outcomes. *Journal of Applied Psychology*, 69, 118-128.
- Dipboye, R. L., Fromkin, H. L., & Wiback, K. (1975). Relative importance of applicant sex, attractiveness, and scholastic standing in evaluation of job applicant resumes. *Journal of Applied Psychology*, 60, 30-43.
- Gorman, C. D., Clover, W. H., & Doherty, M. E. (1978). Can we learn anything about interviewing real people from "interviews" of paper people? Two studies of the external validity of a paradigm. *Organizational Behavior and Human Performance*, 22, 165-192.
- Haefner, J. E. (1977). Race, age, sex, and competence as factors in employee selection of the disadvantaged. *Journal of Applied Psychology*, 62, 199-202.
- Keenan, A. (1977). Some relationships between interviewers' personal feelings about candidates and their general evaluation of them. *Journal of Occupational Psychology*, 50, 275-283.
- Lewis-Beck, M. S., & Mohr, L. B. (1976). Evaluating effects of independent variables. *Political Methodology*, 3, 27-47.
- London, M., & Poplawski, J. R. (1976). Effects of information on stereotype development in performance appraisal and interview contexts. *Journal of Applied Psychology*, 61, 199-205.
- Okanes, M. M., & Tschirgi, H. (1978). Impact of the face-to-face interview on prior judgments of a candidate. *Perceptual and Motor Skills*, 46, 322.
- Pedhazur, E. J. (1982). *Multiple regression in behavioral research* (2nd ed.). New York: Holt, Rinehart & Winston.
- Specht, D. A. (1975). On the evaluation of causal models. *Social Science Research*, 4, 113-133.
- U.S. Department of Labor. (1977). *Dictionary of occupational titles* (4th ed.). Washington DC: U.S. Government Printing Office.

Received July 7, 1986

Revision received April 30, 1987

Accepted May 4, 1987 ■

Call for Nominations for *Journal of Abnormal Psychology*

The Publications and Communications Board has opened nominations for the editorship of the *Journal of Abnormal Psychology* for the years 1990-1995. Don Fowles is the incumbent editor. Candidates must be members of APA and should be available to start receiving manuscripts in early 1989 to prepare for issues published in 1990. Please note that the P&C Board encourages more participation by women and ethnic minority men and women in the publication process, and would particularly welcome such nominees. To nominate candidates, prepare a statement of one page or less in support of each candidate. Submit nominations no later than March 1, 1988 to

Amado Padilla
Department of Psychology
University of California-Los Angeles
Los Angeles, California 90024

Other members of the search committee are Michael J. Goldstein and Anthony J. Marsella.

The Restriction of Range Problem and Nonignorable Selection Processes

Alan L. Gross and Mary Lou McGanney

Department of Educational Psychology, Graduate Center, City University of New York

We describe a general model for estimating test validity when criterion scores cannot be observed for all cases. Unlike the traditional restriction of range correction formula, the proposed procedure can yield accurate estimates in problems in which an unobservable variable, related to y , is the basis for selection. This type of selection process is referred to as being nonignorable. In the nonignorable case, the regression statistics computed in the selected group, as well as the correction formula estimate, will be biased. The potential advantages of the proposed procedure over the correction formula approach are demonstrated in terms of hypothetical data sets in which the selection process is nonignorable. The shortcomings of the proposed analysis and future research needs are also discussed.

The traditional approach to dealing with missing data problems that arise in test validation studies is to use the so-called *restriction of range correction formulas* (Lord & Novick, 1968). In the typical predictive validation study, scores on the test or predictor variable (x) are observed for a sample of N applicants, but criterion or y scores can be observed only for a subset of $N_s < N$ selected subjects. The statistical problem confronting the investigator is to estimate the xy correlation for the applicant population in terms of the available data. The correction formula provides a consistent estimate only if the data satisfy a rather strong set of assumptions. The most crucial assumption concerns the probabilistic relation between the selection process and the y variable; it must be assumed that there is no relation between the potential y score of an applicant and the probability that he or she will be selected, given all other available information. Such a selection process is said to be ignorable (Rubin, 1976). In this special case, one can obtain unbiased estimates of the regression weights and residual variance in predicting y from x simply by analyzing the xy data of the selected group. However, when the selection process is nonignorable (i.e., there is a relation between the probability of selection and y), one cannot extrapolate from the selected group to the full applicant population. The nonignorable selection process produces a pattern of missing data that affects the general shape and orientation of the xy scatterplot in the selected group. The regression weights and residual variance computed in this group will then be biased. Furthermore, because the correction formula is based on these statistics, it too will be biased, typically in the negative direction.

The consequences of a selection process being ignorable or nonignorable can be clearly understood in terms of some simple examples. Consider first the case in which selection is based solely on the x variable; that is, subjects are selected and observed on y if and only if x exceeds some threshold value. Because x is the only basis for selection, an applicant's potential y

score cannot be related to the probability of selection once x is known. Thus, the selection process would be ignorable. This situation is depicted graphically in Figure 1. The scatterplot given in Figure 1A represents the complete xy data set (i.e., the data one would observe if there was no selection). The scatterplot in Figure 1B represents the xy data one would observe when selection is based solely on x . In this ignorable case, it is clear that the selected group estimates of the regression of y on x are representative of the xy regression in the complete group. In addition, the correction formula that is computed in terms of the regression slope and residual variance obtained from the selected group (as well as the variance of x in the total sample) will yield an accurate estimate of the applicant group xy correlation. Another example of an ignorable selection process is given by the scatterplot in Figure 1C. This scatterplot depicts a situation in which selection is based on some unmeasured variable that although correlated with x is unrelated to y , given x . Because x and the selection variable are correlated (suppose positively), the probability of y being observed increases gradually as a function of x . However, because the selection variable is conditionally independent of y , at each given x value the y scores are missing in a random fashion rather than as a function of y . One can see in Figure 1C that the resulting pattern of observed xy data can be analyzed to yield accurate estimates of the applicant group xy relation.

It is not clear if either of the two previously described scenarios are commonplace in real-life validation studies; typically, selection is based on an entire set of variables rather than on x alone, and furthermore, these additional variables cannot always be assumed to be unrelated to y , given x . Consequently, if one only considers x , the probability of selection may be related to y through the relation of y to these unaccounted-for variables. This nonignorable selection process is depicted in Figure 1D. Because the probability of selection and y are related (suppose directly) the y scores are not missing in a random fashion at each x value, as was the case in Figure 1C. Rather, at each x value, the very low y scores tend not to be observed. In the first x array, all five y scores are missing; in the second and third x arrays, the lowest three y scores are missing; and finally, in the fourth array, the lowest y score is missing. It is clear that an analysis of the selected group data in this case will yield under-

Correspondence concerning this article should be addressed to Alan L. Gross, Department of Educational Psychology, Graduate Center, City University of New York, 33 West 42nd Street, New York, New York 10036.

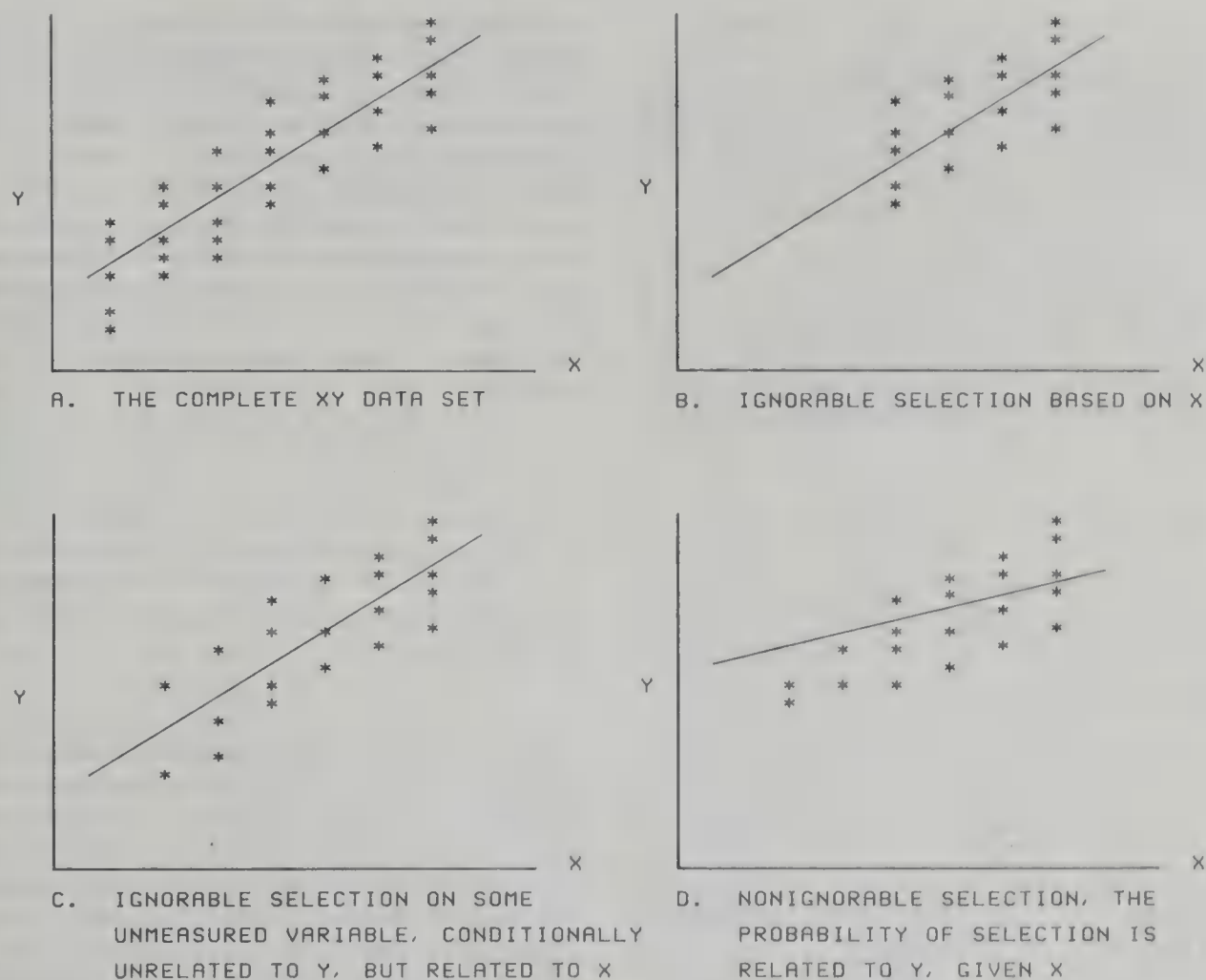


Figure 1. Examples of ignorable and nonignorable selection processes.

estimates of the slope coefficient and residual variance and an overestimate of the regression intercept. Most important, if one were to apply the correction formula in this nonignorable case, the negative bias in the slope coefficient and the residual variance estimate would result in the applicant group xy correlation being underestimated.

Previous research (Gross & Fleischman, 1983; Linn, 1968, 1983; Linn, Harnisch, & Dunbar, 1981) supports the argument that the traditional range restriction formula yields unsatisfactory estimates of the xy correlation when the selection process is nonignorable. Clearly, it would be desirable to have correction procedures that yield consistent estimates of the xy correlation even when the selection process is not ignorable. The basis for such a procedure already exists, having been developed in the field of econometrics (Heckman, 1976, 1979). In the present article we describe how this procedure can be used to estimate the applicant population xy correlation. In the next section the statistical model is described. We then demonstrate the method by analyzing four hypothetical data sets. In the final section the limitations of the proposed procedure are discussed.

The Statistical Model

The proposed approach consists of three components: (a) a regression model that expresses the xy relation, (b) a selection model that describes the selection process (i.e., the basis for the

missing y scores), and (c) an assumption concerning the relation between the two former models. The usefulness of the model stems from the fact that it provides consistent estimates of the xy correlation even when the y variable and the selection process are interrelated. The traditional correction formula approach can be viewed as the special case of the general model in which one assumes *a priori* no interrelationship.

First, consider the regression model that relates y to x in the applicant population. The following model is assumed:

$$y = \beta_0 + \beta_1 x + e_y, \quad (1)$$

where β_0 and β_1 are regression parameters, e_y is the error, and given x , y is normally distributed with mean $\beta_0 + \beta_1 x$, and variance $\sigma^2(y|x)$. The parameter of greatest interest, the xy correlation, can be expressed in terms of the regression parameters:

$$\rho(x, y) = [\beta_1 \cdot \sigma(x)] / [(\beta_1 \cdot \sigma(x))^2 + \sigma^2(y|x)]^{1/2}, \quad (2)$$

where $\sigma(x)$ is the standard deviation of x .

Second, consider the selection model or process that explains the process by which criterion scores (y) can be observed for some applicants, but not for others. Let us assume that selection is based on a variable y_s ; when y_s exceeds some threshold value ($y_s > k$) an applicant is selected and y is observed; otherwise, y cannot be observed. The selection variable y_s is expressed in terms of the following regression model:

$$y_s = \alpha_0 + \underline{\alpha}'\underline{x}_s + e_s, \tag{3}$$

where \underline{x}_s represents a set of $p \geq 1$ observed variables, $\alpha_0, \underline{\alpha}$ are a set of $p + 1$ regression parameters, and e_s is the error variable. The observed selection variables (\underline{x}_s) may in general be one in the same with the predictor variable (x), an entirely different set of variables, or consist of both x and additional variables. It is assumed that given \underline{x}_s , y_s is normally distributed with mean $\alpha_0 + \underline{\alpha}'\underline{x}_s$, and variance $\sigma^2(y_s|\underline{x}_s)$. Note that because the scale of y_s is arbitrary, one can set the variance of y_s equal to 1.0. Furthermore, there is no loss of generality in setting the threshold value k equal to zero. In terms of the model given in Equation 3, the probability that a subject will be selected and observed on y , given \underline{x}_s , is expressed by the normal ogive or probit function:

$$P(\text{select}|\underline{x}_s) = P(y_s > 0|\underline{x}_s) = F(\alpha_0 + \underline{\alpha}'\underline{x}_s), \tag{4}$$

where F is the normal distribution function.

The selection model expressed in Equations 3 and 4 can be explained in the following manner: One can conceptualize the selection process as being based on an entire set of variables, some of which can be observed (\underline{x}_s), and others are unobservable (e_s). The actual selection decision can be viewed as a process whereby a weighted composite of all of the variables is constructed (y_s), and subjects for whom y_s exceeds some threshold value are selected and consequently observed on y . The notion that some of the variables underlying selection are not observable is a realistic one. In cases in which the selection decision is made by the institution, potential selection variables such as letters of recommendation, and performance during an interview may not be formally recorded or quantified. Furthermore, in cases in which self-selection occurs (e.g., individuals drop out), it may be difficult to know the underlying basis for this decision. Thus, in the most general case, the true selection variable (y_s) is unobservable. One can only partially account for this variable in terms of the observed \underline{x}_s variables.

A third component of the model is the assumption concerning the relation between the y and y_s variables, given x and \underline{x}_s . It is assumed that given x and \underline{x}_s , y and y_s are distributed bivariate normal with correlation $\rho(y, y_s|x, \underline{x}_s)$. This correlation is a key parameter of the model because it determines whether the selection process is ignorable (i.e., whether there is a relation between y and the probability of selection). Why is it important to consider this yy_s correlation in estimating the xy correlation? To answer this question, let us consider the regression of y on x in the selected group (Heckman, 1976):

$$\begin{aligned} E(y|x, \text{select}) &= E(y|x, \underline{x}_s, y_s > 0) \\ &= \beta_0 + \beta_1 x + \beta^* x^*, \end{aligned} \tag{5}$$

where β_0, β_1 are the weights appearing in Equation 1,

$$\begin{aligned} \beta^* &= \rho(y, y_s|x, \underline{x}_s) \cdot \sigma(y|x), \\ x^* &= f(\alpha_0 + \underline{\alpha}'\underline{x}_s)/F(\alpha_0 + \underline{\alpha}'\underline{x}_s), \end{aligned}$$

and f and F are the normal density and distribution function, respectively.

Thus, if we consider the prediction of y in the selected group, an additional term based on the x^* variable enters the regression equation. This term reflects the effect of the selection pro-

cess. Note that the term will tend to drop out as the yy_s correlation approaches zero or selection becomes a function of x_s alone. In the former case, $\beta^* = 0$; in the latter case, $x^* = 0$. In the traditional correction formula approach, one operates as if this additional term was not present. Estimates of β_0, β_1 , and $\sigma^2(y|x)$ are then obtained by simply regressing y on x in the selected group. An estimate of the variance of x is readily obtained from the total sample. The actual correction formula estimate $[r_c(x, y)]$ of the population correlation is constructed by considering the expression for the population xy correlation in Equation 2, and replacing the parameters by their sample estimates.

$$\begin{aligned} r_c(x, y) &= [\hat{\beta}_1 \cdot \hat{\sigma}(x)]/[(\hat{\beta}_1 \cdot \hat{\sigma}(x))^2 + \hat{\sigma}^2(y|x)]^{1/2} \\ &= r_{xy_s}/[r_{xy_s}^2 + (s_x^2/S_x^2)(1 - r_{xy_s}^2)]^{1/2}, \end{aligned} \tag{6}$$

where r_{xy_s} = xy correlation in the selected group; s_x^2, S_x^2 = variance of x in the selected and total samples, respectively. $\hat{\beta}_1$ = least squares regression weight computed in the selected group.

$$\hat{\sigma}(x) = S_x.$$

$\hat{\sigma}^2(y|x)$ = residual variance computed in the selected group.

It is clear that this procedure yields consistent estimates only if the selection process is ignorable (i.e., the additional x^* term is not present). However, when this term is present in selected group Equation 5, failure to include it will result in biased estimates for $\beta_0, \beta_1, \sigma^2(y|x)$, and consequently $\rho(x, y)$. When the yy_s correlation is positive, β_0 is overestimated and β_1 and $\sigma^2(y|x)$ are underestimated. Furthermore, the traditional correction formula estimate for the xy correlation tends to be negatively biased.

In the proposed model, one estimates the xy correlation without the overly restrictive assumption of an ignorable selection process. In other words, we need not assume that the β^*x^* term drops out of the equation. At least two different procedures can be used to obtain consistent estimates of the unknown parameters. The first procedure is a two-step method (Heckman, 1976, 1979; Olsen & Becker, 1983). This procedure logically follows from the form of the underlying regression of y on x in the selected group (Equation 5). In the first stage, the α_0 and $\underline{\alpha}$ parameters of the selection model are estimated using a probit analysis in which the probability of selection is predicted from the x_s variables using the entire sample. In the second stage an estimated x^* variable (\hat{x}^*) is constructed for each selected subject as follows:

$$\hat{x}^* = f(\hat{\alpha}_0 + \hat{\alpha}'\underline{x}_s)/F(\hat{\alpha}_0 + \hat{\alpha}'\underline{x}_s), \tag{7}$$

where $\hat{\alpha}_0$ and $\hat{\alpha}'$ are the probit estimates of the selection model parameters, and f and F are the normal density and distribution function, respectively. Using the selected group, the y variable is then regressed on x and \hat{x}^* using ordinary least squares regression. Given estimates of α_0 , and $\underline{\alpha}$ from the first stage, and estimates of $\beta_0, \beta_1, \beta^*$, and $\sigma^2(y|x)$ from the second stage, one can easily obtain estimates for the yy_s and xy correlations by substitution. More specifically, if the estimate for β^* is divided by the estimate for $\sigma(y|x)$, an estimate for the yy_s correlation is obtained. The xy correlation is estimated by replacing the parameters in Equation 2 with their sample estimates. Note

that multicollinearity between the x variable and the constructed x^* variable can lead to inaccurate results when the two-step method is used (Olsen, 1980, 1982). The problem can be best understood by examining Equation 7. It is seen that the \hat{x}^* variable is a transformation of the \underline{x}_s variable. Even though the transformation is nonlinear, if the x and \underline{x}_s variables are highly correlated, the x and x^* variables can also be highly correlated. Clearly, the problem can be at its greatest when x and x_s are the same variables. Multicollinearity can also arise when there is a weak relationship between the y_s variables and the \underline{x}_s variable (i.e., when the selection model parameter α is close to zero). In such a case, the constructed \hat{x}^* variable will be nearly constant for all selected subjects and will, thus, be collinear with the unit constant ($x_0 = 1$) in the regression equation. In general, when either of the two types of multicollinearity are present in the data, the two-step method can produce quite inaccurate results. This problem will be illustrated in the analysis of the hypothetical data sets.

The second estimation procedure is based on the method of maximum likelihood. One estimates the parameters as those values that maximize the joint probability or likelihood of the data. For subjects in the selected group, the data consist of scores on x , \underline{x}_s , and y , and knowledge that $y_s > 0$. For unselected subjects, the data consists of scores on only x and \underline{x}_s , together with the knowledge that $y_s \leq 0$. In many standard statistical problems, the likelihood function can be maximized through the process of first differentiating the logarithm of the likelihood function, setting the resulting derivatives to zero, and then solving for the estimates. However, in the present problem, this process cannot be applied because the equations that must be solved are nonlinear. However, numerical procedures such as the Newton-Raphson (Kendall & Stuart, 1967, pp. 48-49) method can be used to maximize the likelihood function and obtain the corresponding maximum likelihood estimates. Given a set of initial guesses or starting values for the maximum likelihood estimates, the Newton-Raphson method iteratively modifies these starting values until the process converges, hopefully at the maximum of the likelihood function. The starting values for this procedure are typically obtained from the two-step estimates. Although the computation of the maximum likelihood estimates is far more complex than the computation of the two-step estimates, the former estimates have the desirable property of having smaller asymptotic standard errors. One can basically consider the two-step estimates as approximations to the maximum likelihood estimates. Finally, note that although the computations involved in estimating the parameters of the model are not trivial, the entire estimation procedure (both the two-step and maximum likelihood methods) can be easily performed using available computer programs. All of the analysis described in the next section were performed on an IBM compatible personal computer using the program LIMDEP (Greene, 1984). A main frame version of this program is also available.

Sample Analyses

To demonstrate the proposed model, four different hypothetical data sets of size $N = 200$ were analyzed. Each set was computer generated using a Monte Carlo procedure from a specified

population in which the y , y_s , x , and \underline{x}_s variables were jointly normal. The selection models were defined so that the expected size of the selected group would be $N_s = 100$ (i.e., 50% of the y data would be missing). For the first three data sets, the selection process is nonignorable [$\rho(y, y_s|x, x_s) = .70$], whereas for the fourth sample, the selection process was defined to be ignorable [$\rho(y, y_s|x, x_s) = 0.0$].

A typical data set consists of the x and \underline{x}_s scores for all $N = 200$ subjects, the y scores for the N_s selected subjects, and a binary indicator variable (m) having the value of one for a selected subject who is observed on y , or a value of zero for an unselected subject for whom y is missing. The m variable was constructed by first generating the normally distributed latent selection variable (y_s) and then setting $m = 1$ if $y_s > 0$; $m = 0$, if $y_s < 0$. In the first, second, and fourth data sets, a single x variable is used to predict the criterion variable y . However, in the third data set, two x variables are used to predict y . In this case, the population multiple correlation [$\rho(y \cdot x_1, x_2)$], rather than the simple population correlation [$\rho(x, y)$], becomes the validity measure of interest. The population validity of the x variable(s) in predicting y was defined by setting the simple or multiple population xy correlation to a value of .50. Furthermore, the population intercorrelations among the x and x_s variables were set at a common value of .60. Finally, note that the overlap among the x and \underline{x}_s variables is varied over the data sets. In the first data set, the same variable (x_1) is used to predict both y and y_s . However, in the second and fourth data sets different x variables are used for y (x_1) and y_s (x_2). In the third data set, y is predicted from x_1 and x_2 , and y_s is predicted from x_1 .

The analysis of the four data sets is presented in Table 1. The populations from which the data sets were drawn are described in terms of the regression model parameters [$\beta_0, \beta_1, \sigma(y|x)$], the selection model parameters (α_0, α), the coefficient β^* for the additional term that enters the regression equation when y is predicted from x in the selected group (Equation 5), the population correlation between y and the y_s selection variable [$\rho(y, y_s|x, x_s)$], and the population validity of the x variable(s) in predicting y , $\rho(x, y)$, or $\rho(y \cdot x_1, x_2)$. The numerical value of each of these parameters is presented in the second column of Table 1. For each of the four samples, three different analyses are presented. The third column in Table 1 displays the maximum likelihood estimates of the population parameters and their standard errors in parentheses. The two-step estimates are given in the fourth column. Finally, the last column presents the analysis in which β_0, β_1 , and $\sigma(y|x)$ are estimated by simply regressing y on x in the selected group, and the correction formula is used to estimate $\rho(x, y)$. By comparing the results in columns 3, 4, and 5 to the true parameter values given in column 2, one can judge the accuracy of the estimates.

In Data Set 1, the same x variable is used to model both the criterion variable (y) and the selection variable (y_s); that is, $x = x_s$. Furthermore, the selection process is nonignorable because the population yy_s correlation is defined to be .70. First, consider the estimates in column 5, which are obtained by simply predicting y from x in the selected group, and then applying the correction formula given by Equation 6. As expected, the selected group regression equation overestimates the β_0 intercept term and underestimates the regression slope, residual variance, and most important, the population xy correlation.

Table 1
Analysis of Four Hypothetical Data Sets ($N = 200$)

Parameter	Value	Estimate		
		Maximum likelihood	Two step	Selected group
Data Set 1: Nonignorable selection process (both y and y_s are predicted from the same x , $N_s = 110$)				
α_0	.00	.06 (0.10)	.05 (0.10)	
α_1	.58	.61 (0.10)	.62 (0.10)	
β^*	.70		.42 (1.23)	
$\rho(y, y_s x)$.70	.83 (0.15)		
β_0	.00	.01 (0.20)	.26 (0.98)	.59 (0.08)
β_1	.58	.50 (0.13)	.38 (0.42)	.24 (0.08)
$\sigma(y x)$	1.00	.90 (0.12)	.77	.72 (0.05)
$\rho(x, y)$.50	.49	.44	.32
Data Set 2: Nonignorable selection process (y predicted from x_1 , y_s predicted from x_2 , $N_s = 114$)				
α_0	.00	.09 (0.09)	.09 (0.09)	
α_1	.58	.58 (0.11)	.57 (0.11)	
β^*	.70		.83 (0.32)	
$\rho(y, y_s x_1, x_2)$.70	.78 (0.14)		
β_0	.00	-.02 (0.16)	-.10 (0.24)	.46 (0.08)
β_1	.58	.53 (0.10)	.53 (0.09)	.38 (0.07)
$\sigma(y x_1)$	1.00	.90 (0.10)	.95	.75 (0.05)
$\rho(y, x_1)$.50	.51	.49	.45
Data Set 3: Nonignorable selection process (y predicted from x_1 and x_2 , y_s predicted from x_1 , $N_s = 110$)				
α_0	.00	.06 (0.10)	.05 (0.10)	
α_1	.58	.61 (0.10)	.62 (0.11)	
β^*	.70		.41 (1.24)	
$\rho(y, y_s x_1, x_2)$.70	.83 (0.15)		
β_0	.00	-.01 (0.20)	.26 (0.98)	.58 (0.08)
β_1	.32	.23 (0.14)	.12 (0.44)	-.02 (0.10)
β_2	.32	.34 (0.10)	.32 (0.10)	.33 (0.10)
$\sigma(y x_1, x_2)$	1.00	.90 (0.12)	.77	.72 (0.05)
$\rho(y \cdot x_1, x_2)$.50	.48	.45	.38
Data Set 4: Ignorable selection process (y predicted from x_1 , y_s predicted from x_2 , $N_s = 108$)				
α_0	.00	-.12 (0.10)	-.12 (0.10)	
α_1	.58	.62 (0.10)	.62 (0.11)	
β^*	.00		-.43 (0.30)	
$\rho(y, y_s x_1, x_2)$.00	-.33 (0.38)		
β_0	.00	.19 (0.32)	.28 (0.33)	-.05 (0.10)
β_1	.58	.48 (0.11)	.46 (0.13)	.55 (0.10)
$\sigma(y x_1)$	1.00	.96 (0.09)	.98	.94 (0.07)
$\rho(x_1, y)$.50	.45	.42	.51

Note. Standard errors (where available) are given in parentheses.

With respect to this last result, it is seen that although $\rho(x, y) = .50$, the correction formula estimate is only .32. The maximum likelihood estimates in column 3 are clearly more accurate than the selected group estimates. For example, the estimate of the xy correlation now becomes .49. Note that this estimate is obtained simply by considering the expression for $\rho(x, y)$ given in Equation 2, and replacing the parameters β_1 , and $\sigma(y|x)$ by their maximum likelihood estimates, and replacing $\sigma(x)$ by the total sample standard deviation of x . Furthermore, the maximum likelihood estimates of all of the parameters are within one standard error unit of the true values. Finally, it is interesting to note that the two-step estimates given in column 4 are somewhat disappointing when compared to the accuracy of the maximum likelihood estimates. Although more accurate than the

selected group estimates, the two-step procedure still underestimates the regression slope, residual variance, and xy correlation and overestimates the intercept. Note also that the β^* parameter is underestimated (.42) with a rather large standard error (1.23). In general, one can test the hypothesis that a parameter of the model is zero by dividing the estimate by its standard error, and treating the resulting value as a z statistic. If this test was applied to the β^* coefficient in the Data Set 1, one would erroneously conclude that this parameter was not significantly different from zero. In other words, one would incorrectly assume that the selection process is ignorable. On the other hand, the null hypothesis of an ignorable selection process would be rejected if the maximum likelihood results were used. In this case, one would divide the estimate of the yy_s correlation (.83) by its stan-

dard error (0.15). The resulting z value would clearly be significant. The inaccuracy of the two-step method can be attributed to the fact that the x and x_s variables are one and the same in this first data set. Consequently, the correlation in the selected group between the x variable and x^* variable (constructed from x_s) turns out to be quite high, the computed value being $-.98$. A correlation of this magnitude can greatly inflate the standard errors of the estimates of the regression weights.

Data Set 2 differs from the Data Set 1 only in the fact that different x values are used to predict y and y_s . The results again show the advantage of the maximum likelihood method over the selected group analysis. However, the two-step estimates are now almost identical to the maximum likelihood results. The improvement in these estimates, as compared with the first data set, can be attributed to the use of different x variables to predict y and y_s (i.e., a reduction in the correlation between x and x^*). The selected group correlation between x and x^* is now $-.65$ rather than the value of $-.98$ for the first data set. Finally, it is interesting to note that although the selected group method provides rather poor estimates of the regression weights and residual variance, the correction formula estimate (.45) of the xy correlation is negatively biased by only .05 units. This result is due the fact that underestimation of the regression slope is "compensated" for by underestimation of the residual variance.

Data Set 3 involves the use of two x variables to predict y . In this situation, the population validity is expressed by the multiple correlation coefficient. Again, it is clear that the most accurate estimates are obtained from the maximum likelihood method, followed by the two-step method. The traditional analysis, based on using the selected group to predict y from x_1 and x_2 , yields poorer estimates of the regression weights, residual variance, and multiple correlation than do the other two estimation methods. More specifically, it is seen that the selected group estimate of the regression weight for x_1 is $-.02$, and the correction formula estimate of the population validity is only .38. Note that the correction formula given in Equation 6 is easily generalized to the case of more than one x variable by first writing out an expression for the population multiple correlation in terms of the following applicant population parameters: regression weights, residual variance, variances, and covariances among the x s. The correction formula can then be obtained by replacing the population regression weights and residual variance by their estimates obtained in the selected group and replacing the variances and covariances of the x s by the values obtained in the total sample.

Data Set 4 was generated using an ignorable selection process (i.e., the yy_s correlation was set to 0). In this situation the computation of traditional correction formula estimate is the theoretically correct analysis. One can see in the last column of Table 1 that the regression estimates computed in the selected group and the correction formula estimate of the xy correlation are now quite accurate. If one had a priori knowledge that the selection process was ignorable, the best strategy would be to simply use the traditional analysis. Without this a priori knowledge, one would apply the proposed model to the data and estimate the parameter values using either the maximum likelihood or the two-step method. The question of interest in such a case is whether the more general model will yield results that

are similar to those obtained from the traditional analysis. An inspection of the maximum likelihood and two-step estimates supports an affirmative response to the question. First, if one were to test the null hypothesis of an ignorable selection process by testing the significance of the yy_s correlation or the significance of the β^* parameter, it is clear that the resulting z statistics would not be significant. For example, if the estimated yy_s correlation is divided by its standard error, the resulting test statistic is $z = -.33/.38 = -.86$, clearly not significant. Thus, one would correctly infer from Data Set 4 that the selection process is ignorable. Second, an inspection of the actual estimates shows them to be reasonably similar to, but not as accurate as, those obtained from the traditional analysis presented in the last column. This last result can be attributed to the fact that the proposed model requires the simultaneous estimation of more parameters than does the traditional analysis, and thus requires larger sample sizes to achieve reasonably small standard errors.

Summary and Conclusions

In many real-life selection problems, the explicit selection variable cannot be directly observed. Furthermore, this unobserved variable will often be conditionally related to the criterion variable, given the x and x_s variables. In this type of situation, a relation will exist between the probability of selection and the criterion variable, given the observed x and x_s data (i.e., the selection process will be nonignorable). It has been demonstrated both graphically in terms of Figure 1, and algebraically in the Statistical Model section, that a nonignorable selection process produces a pattern of missing y scores that changes the general shape and orientation of the xy scatterplot. Consequently, if one were to attempt to estimate the regression of y on x using the xy data of the selected group, one would typically overestimate the intercept and underestimate the slope and residual variance. Most important, because the correction formula estimate of the applicant group xy correlation is based on the slope and residual variance, it too will be biased, typically in the negative direction. The proposed analysis, on the other hand, can yield improved estimates of the xy correlation as well as the regression parameters, even when the selection process and the y variable are related. In the proposed model, the selection process is modeled in terms of a latent variable (y_s), which may be only partially explainable in terms of observable measures. The possible relation between the criterion variable and the selection process is then accounted for in terms of the correlation between y and y_s . The traditional correction formula approach can be viewed as a special case of the proposed model in which the yy_s correlation is restricted to be zero.

The statistical theory that has been presented, together with the results of the data analysis in the Sample Analyses section, suggest that the proposed analysis can be a useful tool in test validation studies. Should the proposed model become the standard method used by practitioners in investigating the predictive validity of a test? In our opinion, such a recommendation is somewhat premature; there are important questions and problems associated with the proposed analysis that require further investigation.

First, although the proposed analysis does not require the restrictive assumption of an ignorable selection process, it still is

dependent on the assumption of bivariate normality for the joint distribution of the criterion (y) and selection (y_s) variables. How robust is the model with respect to this assumption? This issue has been discussed in some detail by Olsen (1982), where alternative and more general population distributions have been suggested. It would be useful to examine the analysis of real data sets using different distribution assumptions.

Second, the results described in the Sample Analyses section suggest that the analysis requires large sample sizes. For example, in the fourth data set, in which the selection process was defined to be ignorable, the maximum likelihood and two-step estimates were not as accurate as the selected group estimates using a total sample size of $N = 200$ and a selected group size of $N_s = 110$. It would be useful to have some guidelines as to the necessary sample sizes.

Third, the numerical computation of the estimates is not at all straightforward. With respect to the two-step estimates, the problem of multicollinearity can be encountered if (a) the x and x_s variables are highly correlated (the worst case being one in which $x = x_s$; i.e., the same variable is used to predict both y and y_s), or (b) the y_s variable is poorly predicted from the x_s variables (i.e., the probit analysis yields near zero estimates for the α parameter). In either case, the collinearity can result in estimates having unacceptably large standard errors. Furthermore, because the two-step estimates are used as starting values for the iterative computation of the maximum likelihood estimates, the iterative procedure may not converge, or may converge, to incorrect estimates if the starting values are poor. Olsen (1982) suggested a conditional maximization procedure that can overcome the problem of poor starting values. Unfortunately, this procedure is not currently available in the LIMDEP program. Clearly, it would be useful to investigate these computational problems by conducting analyses of additional data sets.

Fourth, the proposed analysis assumes that the selection process depends on whether a latent y_s variable exceeds some threshold value. This conceptualization may not be appropriate for all test validation studies. For example, Linn (1983) described a selection process at a junior college, where criterion scores are missing not only for applicants who are rejected by the institution but also for those who decline an admission offer and enter a 4-year college. Thus, the individuals for whom y scores can be observed may represent some middle-ability-level group, the low- and high-ability cases being removed by institutional and self-selection, respectively. Such a selection process may be more realistically modeled in terms of whether a latent selection variable falls in some intermediate range, ($k_1 < y_s < k_2$) rather than simply exceeding a single threshold value. Thus, some modifications of the proposed model may be necessary in particular validation studies.

In conclusion, the proposed model represents an improvement over the traditional correction formula approach to estimating test validity when criterion scores cannot be observed for all applicants. However, more research is needed with respect to the questions previously raised. In the past, the development of methods for dealing with nonignorable selection processes has been conducted primarily in the field of econometrics (Maddala, 1983). The importance of this problem to the area of test validation studies, hopefully, will stimulate research by psychometricians.

References

- Greene, W. H. (1984). LIMDEP [Computer program]. New York: New York University, Department of Economics, Graduate School of Business Administration.
- Gross, A. L., & Fleischman, L. (1983). Restriction of range corrections when both distribution and selection assumptions are violated. *Applied Psychological Measurement*, 2, 227-237.
- Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection, and limited dependent variables, and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5, 475-492.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153-161.
- Kendall, M. G., & Stuart, A. (1967). *The advanced theory of statistics* (Vol. 2). New York: Hafner Publishing.
- Linn, R. L. (1968). Range restriction problems in the use of self-selected groups for test validation. *Psychological Bulletin*, 69, 69-73.
- Linn, R. L. (1983). Pearson correction formulas: Implications for studies of predictive bias and estimates of educational effects in selected samples. *Journal of Educational Measurement*, 20, 1-15.
- Linn, R. L., Harnisch, D. L., & Dunbar, S. B. (1981). Corrections for range restriction: An empirical investigation of conditions resulting in conservative corrections. *Journal of Applied Psychology*, 66, 655-663.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Maddala, G. S. (1983). *Limited dependent and qualitative variables in econometrics*. New York: Cambridge University Press.
- Olsen, R. J. (1980). A least squares correction for selectivity bias. *Econometrika*, 48, 1815-1820.
- Olsen, R. J. (1982). Distributional tests for selectivity bias and a more robust likelihood estimator. *International Economic Review*, 23, 223-240.
- Olson, C. A., & Becker, B. E. (1983). A proposed technique for the treatment of restriction of range in selection validation. *Psychological Bulletin*, 93, 137-148.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Received July 21, 1986

Revision received March 16, 1987

Accepted May 4, 1987 ■

Detecting Infrequent Deception

Kevin R. Murphy
Colorado State University

Recent proposals for using the polygraph and similar devices in routine screening have been aimed at detecting deception in situations sometimes characterized by low base rates. Equations are developed that show that extraordinarily high levels of accuracy would be needed to detect infrequent deception. In this context, the debate over the accuracy of these methods is irrelevant; the accuracy needed to detect infrequent deception far exceeds the levels claimed by the most enthusiastic proponents of these detection techniques. The limits on the use of any particular test of deception can be determined by considering the base rate for deception and the proportion of the nondeceptive population that fails the test. When the base rate is less than .10, these limits are extremely restrictive.

Methods of detecting deception, especially the polygraph, have been the subject of a great deal of research and debate, much of which has centered on the possibility that innocent parties will be wrongly labeled as deceptive (Lieblich, Ben-Shakhar, Kugelmass, & Cohen, 1978; Lykken, 1974, 1978, 1979, 1981, 1985; Podlensky & Raskin, 1977; Raskin & Podlensky, 1979; Sackett & Decker, 1979). Recently the American Psychological Association Council of Representatives passed a resolution on the polygraph that stated in part that

There is the possibility of great damage to innocent persons who must inevitably be labeled as deceptors in situations where the base rate of deception is low; an unacceptable number of false positives would occur even should the validity of testing procedures be quite high. ("Council Takes Stand," 1986)

Polygraphs and similar methods are used in several contexts in which the base rate for deception is likely to be low. For example, nearly three fourths of all polygraph examinations are conducted for the purpose of preemployment screening (Kleinmuntz, 1985) and often focus on employee theft. The base rate for nontrivial employee theft (e.g., more than \$5) is less than 5% in most settings (Hollinger & Clark, 1983), and the base rate for other types of nontrivial deception (except for questions involving drug use) in this context is thought to be equally low (U.S. Office of Technology Assessment, 1983).

It has long been known that base rates have a substantial impact on the accuracy of tests of every sort and that accurate detection of any condition (e.g., deception, psychopathology) for which the base rate is low is difficult (Dawes, 1962; Lykken, 1974; Meehl & Rosen, 1955; U.S. Office of Technology Assessment, 1983). The dual purpose of this article is to demonstrate the boundary conditions for detecting deception and to show that the accurate detection of infrequent deception demands impossibly high levels of validity.

I thank Jeanette Cleveland, George Thornton, and two anonymous reviewers for their valuable comments.

Correspondence concerning this article should be addressed to Kevin R. Murphy, Department of Psychology, Colorado State University, Fort Collins, Colorado 80523.

Formulation

The problem of detecting deception can be analyzed in terms of the base rate and the conditional probability of deception, given the results of a test of deception (Brett, Phillips, & Beary, 1986; Meehl & Rosen, 1955). Here, the base rate defines the prior probability of guilt, or deception. Thus, when 5% of all subjects are involved in illegal activities that they wish to conceal, the probability that a subject selected at random is so involved is .05. A polygraph, an honest test, a voice analysis, or some other method of detecting deception provides data that may lead to a revised estimate of the likelihood of deception. Finally, there must be some threshold for determining when the probability of deception is sufficiently high to conclude that the subject is, in fact, attempting to deceive the questioner.

One of the most useful equations for analyzing the problem at hand is

$$\frac{P(D|F)}{P(T|F)} = \frac{P(F|D)}{P(F|T)} \times \frac{P(D)}{P(T)}, \quad (1)$$

where D is the hypothesis that the subject is deceptive, T indicates that the subject is telling the truth, F indicates that the subject fails the test and is labeled deceptive, $P(D)$ is the unconditional probability or base rate for deception, $P(D|F)$ is the conditional probability that the subject is deceptive, given the fact that the subject has failed the test, and $P(F|T)$ is the false positive rate (i.e., the proportion of nondeceptive subjects who failed the test).

Equation 1, which can be derived from the basic definition of conditional probability (c.f. Dawes, 1962), defines the odds form of Bayes Theorem. Equation 1 is equivalent to the statement *Posterior Odds = Likelihood Ratio \times Prior Odds*. Here, the prior odds refer to the base rate, the likelihood ratio refers to the sensitivity of the test in discriminating deceptive from truthful subjects, and the posterior odds refer to the assessment, after the test, of the likelihood of deception. Prior and posterior odds are defined as the ratio of the probability that a person is deceptive to the probability that a person is not deceptive. When the odds are less than 1.0 (one to one), the evidence favors the hypothesis that the subject is not deceptive; odds greater than

Table 1
Maximum False Positive Rates That Would Allow Investigator To Meet Minimum Threshold for Deception

Base rate for deception	Maximum permissible false positive rate
.30	.428
.20	.250
.10	.111
.05	.052
.03	.030
.01	.010

1.0 indicate that the evidence favors the hypothesis that the subject is deceptive.

It is possible to use Equation 1 to define a minimum threshold for deciding whether a subject is deceptive, which, in turn, can be used to establish the boundary conditions for detecting infrequent deception. This formulation also provides simple methods of incorporating factors such as reasonable doubt and the possibility of false negatives in defining these boundary conditions.

Minimum Threshold

The minimum threshold for labeling a subject deceptive occurs when the data in favor of the hypothesis *deceptive* are at least as strong as the evidence against this hypothesis, which occurs when the posterior odds of deception are equal to 1.0. When the posterior odds are less than 1.0, the data favor the hypothesis *not deceptive*; it would be irrational to go against the weight of the available evidence and label a subject deceptive unless the potential damage associated with deception were very great (Ben-Shakhar, Lieblich, & Bar-Hillel, 1982; Lieblich, et al., 1978). As can be seen from Equation 1, posterior odds of 1.0 are attained when the likelihood ratio for a test is the reciprocal of the prior odds. For example, if the base rate for nontrivial employee theft were assumed to be .05, the prior odds that an employee chosen at random is a thief would be .052 (i.e., .05/.95), or 1 in 19. A test of deception would therefore have to have a likelihood ratio of 19 to provide any chance of concluding that this subject was deceptive. In concrete terms, this means that the true positive rate— $P(F|D)$ —would have to be 19 times as large as the false positive rate— $P(F|T)$. If the likelihood ratio were less than 19, the test could not possibly lead to a situation in which the preponderance of evidence favored guilt. Thus, even if a test were perfectly accurate in detecting actual deception, a test with a false positive rate greater than .052 would not be sufficiently sensitive to detect deception that occurred in 5 cases out of 100.

Note that the conditional probabilities referred to above— $P(F|D)$ and $P(F|T)$ —describe the true positive and false positive rates, or the proportions of deceptive and nondeceptive subjects who fail the test. These rates are always greater than or equal to the unconditional probabilities of a true positive and a true negative outcome. In particular, the unconditional probability of a false positive is given by $P(FP) = P(T) \times P(F|T)$. Thus, if the false positive rate for a test is .052, and 90% of those

who take the test are telling the truth, the probability of a false positive outcome is .047.

If posterior odds of 1.0 are accepted as a lower bound for labeling a subject deceptive, the false positive rate of a test must be very low to allow any possibility of detecting infrequent deception. Assume, for example, that the test catches every person who attempts deception— $P(F|D) = 1.0$. In this case, the probability that a nondeceptive subject fails the test— $P(F|T)$ —must be less than or equal to the prior odds of deception; otherwise, the test will never provide evidence that favors the verdict *guilty*.

Table 1 lists the maximum false positive rates— $P(F|T)$ —that could be tolerated given this minimum threshold for guilt. The table suggests that when base rates are low (e.g., .10 or less) this maximum is approximately equal to the base rate. That is, when deception is infrequent, the tendency of the test to label innocent subjects as deceptive must be equally infrequent.

Reasonable Doubt

Assume that the posterior odds were 100 to 99 (i.e., odds of 1.01) that the subject was attempting to deceive. Although the weight of the available evidence favors a guilty verdict, the evidence is not very strong and would not be persuasive in most settings. The concept of *reasonable doubt* implies that the evidence in favor of guilt (deceptive) must, in most cases, strongly outweigh the evidence in favor of innocence (not deceptive) before a verdict of *guilty* will be pronounced. That is, the hypothesis that a subject is deceptive will not ordinarily be accepted unless the posterior odds of deception are considerably larger than one to one.

Simon and Mahan's (1971) survey of judges, juries, and lay people suggests that a legal criterion of reasonable doubt would translate into posterior probabilities of guilt between .85 and .90. Conventions for testing statistical hypotheses suggest that psychologists employ values of .95 or .99 in quantifying reasonable doubt. For the purpose of illustration, assume the criterion that a posterior probability of guilt of .90 or greater is sufficient to label a subject *deceptive*. This translates into posterior odds of 9 to 1.

Table 2 illustrates the maximum false positive rates— $P(F|T)$ —that could be tolerated given this threshold for guilt. The values in Table 2 are obtained by dividing the corresponding values in Table 1 by 9, reflecting the fact that this 90% threshold for guilt requires posterior odds 9 times as large as the break-even threshold of one to one.

Table 2
Maximum False Positive Rates That Would Allow Investigator To Reach 90% Confidence Threshold for Deception

Base rate for deception	Maximum permissible false positive rate
.30	.047
.20	.027
.10	.012
.05	.005
.03	.003
.01	.001

Table 3
Maximum Permissible False Positive Rates
if Test Can Be Deceived

Base rate	PDSC .98		PDSC .95		PDSC .90		PDSC .85	
	1:1	9:1	1:1	9:1	1:1	9:1	1:1	9:1
.30	.419	.046	.406	.044	.385	.042	.363	.034
.20	.245	.026	.237	.025	.225	.024	.212	.022
.10	.108	.011	.105	.014	.099	.010	.094	.010
.05	.050	.004	.049	.004	.046	.004	.044	.004
.03	.029	.002	.028	.002	.027	.002	.025	.002
.01	.009	.0009	.009	.0009	.009	.0009	.009	.0009

Note. PDSC = Proportion of deceptive subjects caught by test. Ratios represent posterior odds needed to conclude subject is deceptive.

The 90% threshold is conservative in that it implies a strong bias toward the verdict of *not guilty*. This bias may be appropriate if the test of deception leads directly to potentially adverse actions, as is frequently the case in preemployment screening. If a test of deception does not lead directly to any action, a less conservative threshold might be in order. Maximum false positive rates can be determined for any other guilt threshold by putting that threshold in odds form (e.g., an 80% confidence threshold equals odds of 4 to 1 that the person who fails the test is in fact deceptive) and dividing the values in Table 1 by those odds.

Deceiving Tests of Deception

Tables 1 and 2 are based on the assumption that tests of deception are always successful in detecting attempted deception—an assumption that is not supported by empirical research (Lykken, 1979; Sackett & Harris, 1984). When the probability of successful deception is taken into account, the limits placed by false positive rates on the detection of infrequent deception become even more extreme than those portrayed in Tables 1 and 2. Maximum permissible false positive rates, listed as a function of the base rate, the threshold for labeling a subject *deceptive*, and the probability that a deceptive individual will be caught, are shown in Table 3. The values that appear there are obtained by multiplying the corresponding values in Tables 1 and 2 by the probability of successfully detecting attempted deception— $P(F|D)$.

Several features of Table 3 are noteworthy. First, when the base rate for deception is low, the false positive rate must also be low, regardless of the threshold employed. Second, and more important, at very low base rates (e.g., .10 or lower), variations in the ability of the test to detect true deception have little impact on the overall worth of the test. For example, if the base rate for deception were .01, raising the test's probability of detecting actual deception from .85 to .98 (this would be an impressive improvement!) would have no discernible impact on the proportion of false positives that could be tolerated. If the base rate were .01, and more than 9 out of every 1,000 nondeceptive subjects failed the test, the test could not provide a preponderance of evidence to support the hypothesis that any individual examinee was deceptive. If reasonable doubt is factored in, this figure will be even lower.

Conclusion

The use of the polygraph and other similar devices for detecting and deterring deception has been the subject of intense debate in recent years. The use of these devices for screening purposes (e.g., preemployment screening) is of special concern, because in these settings the base rate for deception may be quite low. When the base rate for nontrivial deception is low (e.g., less than .10), tests must be extraordinarily sensitive to provide convincing evidence of deception. For example, if a polygraph were accurate in 98% of the cases examined (this figure is higher than the level of accuracy claimed by proponents of the polygraph, and considerably higher than the accuracy levels shown in empirical research), and errors were equally divided between false positives and false negatives, the false positive rate would be at least .0101, effectively ruling out the polygraph for base rates of .01 or lower. If a 90% reasonable doubt criterion were applied, a test with a 98% accuracy level would be too inaccurate to detect deception when the base rate is as high as .09.

It is important to note that the analyses presented here apply to any technique that might be used to detect a condition for which the base rate is low (Meehl & Rosen, 1955). For example, the Department of Defense has tested over 600,000 individuals for acquired immune deficiency syndrome (AIDS), and approximately 1.5 cases per 1,000 have tested positive ("AIDS testing screens 287," 1986). If the base rate for AIDS exposure is approximately .0015, a false positive rate as low as 2 per 1,000 would be sufficiently high to call into question the AIDS screening program.

It is not the purpose of this article to argue that tests of deception *are* useful when the base rate for deception is high; their validity is doubtful regardless of the base rate. In addition, serious ethical questions have arisen, particularly the possibility that the polygraph represents a "psychological fourth degree" that is used to extract confessions from suspects (Furedy, 1985; Furedy & Liss, 1985). Thus, even if the polygraph *were* valid, its use would be problematic. However, when the base rate is low, this analysis suggests that the debate over the value of the polygraph is meaningless because the accurate detection of deception may be beyond the capability of *any* known test or procedure.

References

- AIDS testing screens 287 from military. (1986, December 13). *The Denver Post*, p. 5A.
- Ben-Shakhar, G., Liebllich, I., & Bar-Hillel, M. (1982). An evaluation of polygraphers' judgments: A review from a decision theoretic perspective. *Journal of Applied Psychology*, 67, 701-713.
- Brett, A. S., Phillips, M., & Beary, J. F. (1986, March 8). Predictive power of the polygraph: Can the "lie detector" really detect liars? *Lancet*, pp. 544-547.
- Council takes stand on AIDS, polygraph. (1986, March). *APA Monitor*, p. 11.
- Dawes, R. M., (1962). A note on base rates and psychometric efficiency. *Journal of Consulting Psychology*, 26, 422-424.
- Furedy, J. J. (1985, January). Some post-Phillonic flights of polygraphic fancy. *Criminal Lawyers' Association Newsletter*, p. 3.
- Furedy, J. J., & Liss, J. (1985). Countering confessions induced by the

- polygraph: Of confessionals and psychological rubber hoses. *Criminal Law Quarterly*, 29, 91-114.
- Hollinger, R. D., & Clark, J. P. (1983). *Theft by employees*. Lexington, MA: Lexington Books.
- Kleinmuntz, B. (1985). Lie detectors fail the truth test. *Harvard Business Review*, 85, 36-42.
- Lieblich, I., Ben-Shakhar, G., Kugelmass, S., & Cohen, Y. (1978). Decision theory approach to the problem of polygraph interpretation. *Journal of Applied Psychology*, 63, 489-498.
- Lykken, D. (1974). Psychology and the lie detector industry. *American Psychologist*, 29, 725-739.
- Lykken, D. (1978). Uses and abuses of the polygraph. In H. L. Pick (Ed.), *Psychology from research to practice* (pp. 171-191). New York: Plenum Press.
- Lykken, D. (1979). The detection of deception. *Psychological Bulletin*, 86, 47-53.
- Lykken, D. (1981). *A tremor in the blood*. New York: McGraw-Hill.
- Lykken, D. (1985). The case against the polygraph in employment screening. *Personnel Administration*, 30, 59-65.
- Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, 52, 194-216.
- Podlensky, J. A., & Raskin, D. C. (1977). Physiological measures and the detection of deception. *Psychological Bulletin*, 84, 782-799.
- Raskin, D. C., & Podlensky, J. A. (1979). Truth and deception: A reply to Lykken. *Psychological Bulletin*, 86, 54-59.
- Sackett, P. R., & Decker, P. J. (1979). Detection of deception in the employment context: A review and critical analysis. *Personnel Psychology*, 32, 487-506.
- Sackett, P. R., & Harris, M. M. (1984). Honesty testing for personnel selection. *Personnel Psychology*, 37, 221-245.
- Simon, R. J., & Mahan, L. (1971). Quantifying burdens of proof: A view from the bench, jury, and classroom. *Law and Society Review*, 5, 319-330.
- U.S. Office of Technology Assessment. (1983). *Scientific validity of polygraph testing: A research review and evaluation* (Tech. Mem. OTA-TM-H-IS). Washington, DC: U.S. Government Printing Office.

Received August 5, 1986

Revision received March 23, 1987

Accepted March 26, 1987 ■

Prosocial Behavior, Noncompliant Behavior, and Work Performance Among Commission Salespeople

Sheila M. Puffer
State University of New York at Buffalo

In this study I identified two types of nontask behavior, prosocial and noncompliant, and tested some of their antecedents as well as their relation to work outcomes. Prosocial behavior represented nontask behaviors that benefited the organization, and noncompliant behavior represented nontask behaviors that were dysfunctional to the organization, as rated by supervisors. Need for achievement, satisfaction with material rewards, and low perceived peer competition were related to prosocial behavior. Low need for achievement and low confidence in management were related to noncompliant behavior. Noncompliant behavior was negatively associated with performance, but prosocial behavior was nonsignificant when noncompliant behavior was controlled. A dual-factor theory (Herzberg, 1966) of nontask behavior is proposed.

An implicit assumption in the study of work performance has been that performance outcomes are dependent on role behavior associated with specific tasks and are governed by organizational appraisal and reward systems (see Iaffaldano & Muchinsky, 1985, and Schneider, 1985, for examples and reviews). Yet, a work role encompasses a diversity of behavior. Focal tasks are the most clearly specified and most readily measured by productivity standards of quantity and quality (e. g., number of units produced). No less important to organizations are nontask behaviors that are relevant to the work context but not directly related to focal tasks. Nontask behaviors that have positive implications for organizations (e. g., volunteering to help a new employee) are referred to as *prosocial behavior*. Nontask behaviors that have negative organizational implications (e. g., those that present a negative image of the organization) are referred to as *noncompliant behaviors*. Whether nontasks are formalized in job descriptions and performance appraisals is not as significant as the fact that nontask behaviors are more difficult to control, reward, and motivate than are focal task behaviors (Brief & Motowidlo, 1986). Nontask behaviors may stem from different motivational bases (Katz, 1964) and situational contingencies than do focal task behaviors in that performing them is more at the discretion of the incumbent. It is conceivable that an individual whose focal task performance was highly valued would be retained in spite of not performing desired

prosocial activities. Similarly, a top performer might be retained in spite of noncompliant behavior.

This article examines nontask behaviors in two ways. First, it tests whether prosocial and noncompliant behavior have different relations to three constructs—achievement and autonomy motives, personal security, and reciprocity. Second, it explores the relation that prosocial and noncompliant behavior have with work performance, which denotes outcomes produced in a job from task and nontask behaviors. A related issue is the extent to which independent variables are correlated with work performance, with prosocial and noncompliant behavior partialled out.

Two Sides of Nontask Behavior

Three recent studies of citizenship behavior were the impetus for this formulation of nontask behavior. C. Smith, Organ, and Near (1983, p. 657) identified two types of citizenship behavior in a 16-item factor analysis: altruism (e. g., "helping a specific person in face-to-face situations") and generalized compliance (e. g., "compliance with internalized norms defining what a 'good employee ought to do'"). O'Reilly and Chatman (1986, p. 495) adapted these items and labeled the resulting factors *extrarole* ("actions for which the individual receives no immediate reward and which benefit the larger organization") and *intrarole* ("behaviors required by the job description"). Williams, Podsakoff, and Huber (1986) found that the C. Smith et al. (1983) citizenship scale produced three factors: altruism, impersonal conscientiousness, and attendance. These studies show considerable ambiguity surrounding the construct of citizenship behavior. The present study proposes prosocial and noncompliant behavior as a means of interpreting the findings in these studies.

Prosocial Behavior

According to Brief and Motowidlo (1986, p. 711), prosocial organizational behavior is

This article is based on my doctoral dissertation at the University of California, Berkeley. Support for this research was provided by the Social Sciences and Humanities Research Council of Canada, the State University of New York at Buffalo, and the Retail Management Institute at the University of Santa Clara.

I am indebted to my committee members, Barry M. Staw, Charles A. O'Reilly, and Philip E. Tetlock, for their guidance. The assistance of David F. Caldwell, Jill W. Graham, Janet P. Near, and James H. Fraser is gratefully acknowledged.

Correspondence concerning this article should be addressed to Sheila M. Puffer, Jacobs Management Center, State University of New York, Buffalo, New York 14260.

behavior which is (a) performed by a member of an organization, (b) directed toward an individual, group, or organization with whom he or she interacts while carrying out his or her organizational role, and (c) performed with the intention of promoting the welfare of the individual, group, or organization toward which it is directed.

This definition is adopted here, except that prosocial behavior is restricted to behaviors compatible with organizational goals.

Numerous types of prosocial organizational behavior have been studied as altruism (e. g., C. Smith et al., 1983) and helping behavior (e. g., Dovidio, 1984). Interpersonal helping has included subordinates helping a supervisor (e. g., Berkowitz & Daniels, 1963; C. Smith et al., 1983), supervisors helping subordinates (Burke & Weir, 1978; Burke, Weir, & Duncan, 1976; Kaplan & Cowen, 1981), and coworkers helping each other (Burke, 1982a, 1982b; C. Smith et al., 1983). Behaviors that benefit the organization as a whole have included voluntary attendance at work during a snowstorm (F. Smith, 1977), performing work-related duties in off-hours (Wiener & Gechman, 1977), housecleaning (Bateman & Organ, 1983), attending company functions (C. Smith et al., 1983), planning social events, and donating money (O'Reilly & Chatman, 1986).

Noncompliant Behavior

Noncompliant behavior refers to breaking rules or norms and describes behaviors that have been previously incorporated in generalized compliance, intrarole behavior, impersonal conscientiousness, and attendance. It is argued that the act of not complying with established rules and practices is more informative about an individual than is compliance. There may be many reasons why people obey the rules—to fulfill the psychological contract of standard business conduct, to avoid sanctions, and because of habit, to name a few. Taken to its extreme, compliance can be phony and even destructive (Bunker, Banckert, Di Biase, & Mc Gillicuddy, 1986). Because compliance is the norm and an expected part of the job, it seems inappropriate to classify it as prosocial behavior. This may have been O'Reilly and Chatman's (1986) implicit rationale for denoting such behavior as intrarole. However, noncompliant behavior can be construed as a separate category in that it is probably more active, deliberate, and premeditated than simply obeying the rules (assuming that the rules are known and understood). Williams et al.'s (1986) observation that their impersonal conscientiousness factor had only negatively worded items is consistent with the notion that the violation of rules is more significant than their observance. Noncompliant behavior is also related to the literature on sanctions and discipline that studies ways of managing noncompliant or undesired behavior (e. g., Beyer & Trice, 1984; O'Reilly & Weitz, 1980; Podsakoff & Todor, 1985; Podsakoff, Todor, & Skov, 1982).

Hypotheses

Three Potential Explanations of Prosocial and Noncompliant Behavior

Achievement and autonomy motives. The need for achievement refers to the need to strive to do things as well and as quickly as possible (Murray, 1938). Worthy (1986, p. 1) ad-

vanced the thesis that "human beings have an inherent tendency to overachieve at work rather than underachieve" because pride in one's work is a reflection of pride in oneself. He viewed prosocial behavior as a means of expressing the need for achievement, inasmuch as prosocial behavior represents extra effort and conscientiousness. Accordingly, a positive relation is predicted between need for achievement and prosocial behavior, and a negative relation between need for achievement and noncompliant behavior.

Murray's (1938) construct of the need for autonomy includes defiance of authority and convention, independence, and irresponsibility. People with a high need for autonomy may pay little attention to opportunities for engaging in prosocial behavior, which suggests a negative relation between autonomy and prosocial behavior. On the other hand, a high need for autonomy may be positively linked to noncompliant behavior in that people may actualize their need to defy authority and convention by violating rules.

Personal security. Studies of interpersonal helping have shown that people who feel deprived or unfairly treated are less inclined to help others. For example, subjects who felt underpaid declined more frequently to donate to a needy family than did subjects who felt equitably paid (Miller, 1977); similarly, subjects induced to imagine their best friend on a holiday were less helpful than those who imagined themselves on the holiday (Rosenhan, Salovey, & Hargis, 1981); and employees low in job satisfaction exhibited low altruistic behavior (C. Smith et al., 1983) and low consideration for others (Motowidlo, 1984). These findings suggest that, consistent with self-awareness theory (Duval & Wicklund, 1972), insecurity about one's personal situation impedes one's ability to focus on and react to external situations.

Two variables reflecting personal security are satisfaction with material rewards and perceived peer competition. By satisfying material needs, financial and job security are expected to reduce preoccupation with one's personal situation and facilitate prosocial behavior. However, the perception of competition from coworkers in attaining performance goals may be personally threatening to the extent that the superior performance of others threatens one's earnings or job security. Perceived peer competition could therefore lead to focusing on oneself, to the exclusion of the external environment wherein opportunities exist for performing prosocial behaviors. No relation is expected, however, between personal security and noncompliant behavior. For example, people risk being punished, losing wages or their jobs. Furthermore, a sense of professionalism or pride may restrain people from acting on their dissatisfaction or insecurity.

Reciprocity. The norm of reciprocity holds that people help those who have helped them (Gouldner, 1960). For example, Berkowitz and Daniels (1964) attributed subordinates helping their supervisors to reciprocity. Eisenberger, Huntington, Hutchison, and Sowa (1986) found that absenteeism was lower among employees who perceived that the organization valued their contributions and cared about their well-being. The notion of fairness underlies the norm of reciprocity in that people seek to balance the inputs and outcomes they have in relation to others (Dovidio, 1984). Faith in peers denotes the belief that one's peers can be trusted to do such things as helping others if

there is a need. Confidence in management refers to trust established through leader supportiveness and disciplinary style. These two variables are expected to be positively related to prosocial behavior and negatively related to noncompliant behavior.

Relation to Work Performance

In the foregoing hypotheses, prosocial and noncompliant behavior are treated as dependent variables. They are now considered to be independent variables in relation to work performance. With respect to prosocial behavior, Katz (1964, p. 132) pointed out that dependable performance of one's prescribed role is no guarantee of organizational effectiveness. It must be supplemented by innovative and spontaneous behaviors initiated by organizational members in reaction to unanticipated events. Cooperative behavior has also been recognized as an important component of productivity at both the micro and macro levels of analysis (Weiss, 1984, pp. 165-168). Prosocial behavior may be positively related to work performance for two reasons. First, people may believe, rightly or wrongly, that the extra behaviors indirectly improve their work performance, thereby increasing their rewards. Second, even if individuals view prosocial behavior and work performance as unrelated, individual characteristics (such as need for achievement) may lead them to perform focal tasks and prosocial behaviors equally well. Alternatively, prosocial behavior and performance may be unrelated if the benefits of prosocial behavior are restricted to intangibles such as company reputation or goodwill.

A negative relation is predicted between noncompliant behavior and work performance. Presumably, work rules are imposed with the aim of encouraging high work performance, hence noncompliance should lead to low performance. Several studies have found that certain leader behaviors are effective in reducing undesired behavior and thereby increasing performance (e. g., Beyer & Trice, 1984; O'Reilly & Weitz, 1980; Podsakoff et al., 1982).

Method

Sample

This study was conducted in 12 stores of a California retail furniture chain. All salespeople, paid strictly on commission and selling the same line of merchandise, were surveyed on a voluntary basis. Most salespeople completed the questionnaire at a regularly scheduled staff meeting prior to their shift. Response rate was 90% (141 out of 157 questionnaires distributed); six nonrespondents were known to be on vacation, and two others were newly hired and declined due to insufficient familiarity with the organization. Mean experience in commission sales with the company was 9.2 years, and total experience in sales averaged 19.8 years. In all, 71% of the respondents were men. A total of 17% had a high school diploma; 53% had some college education; and 30% had a college degree.

Procedure

The data were collected from three sources: Sales performance was obtained from company records; store managers provided assessments of salespeople's prosocial and noncompliant behaviors; and salespeople

Table 1
Measures and Factor Loadings for Prosocial and Noncompliant Behaviors

No.	Item	Factor loading
Prosocial behavior		
1.	Investigating postsale customer service problems	.65
2.	Assisting other salespeople (e.g. handling their customers' postsale problems)	.84
3.	Housekeeping (e.g., keeping product displays and catalogs tidy)	.76
4.	Marking the floor for special sales (i.e., attaching sales tags to merchandise)	.70
5.	Assisting in general store operations (e.g., opening and closing the store)	.87
Eigenvalue		2.97
% variance explained		.59
Noncompliant behavior		
1.	Being late, taking excessive breaks	.81
2.	Complaining about the company or coworkers	.76
3.	Violating floor rules (e.g., taking sales from other salespeople)	.65
4.	Making unrealistic promises to customers (e.g., promising early delivery dates)	.71
5.	Failing to do fair share of noncommission sales promotions (e.g., phoning customers about upcoming sales; selling account insurance)	.77
Eigenvalue		2.75
% variance explained		.55

completed a questionnaire containing the independent variables and demographic information.

Prosocial and noncompliant behaviors. Each of the 12 store managers provided a list of prosocial behaviors and noncompliant behaviors that they had observed among salespeople. Managers rated on 7-point scales how important each behavior was and how frequently the opportunity arose to perform it. The 5 most important and frequent prosocial behaviors common to all stores were retained, as were the 5 most important and frequent noncompliant behaviors. Managers then rated their salespeople on how frequently they had observed them perform each of the 10 behaviors in the past year (1 = *never*, 7 = *always*). (See Table 1.)

The corporate management explained that although they used sales performance as the criterion to evaluate and retain sales personnel, they also valued behaviors, above and beyond the sales role, that helped the company but that were related only indirectly to sales. For instance, management encouraged salespeople to handle postsale customer service problems, but they neither sanctioned those who failed to do so, nor compensated those who did; administrative staff handled customer problems if salespeople did not. Management believed that salespeople had sufficient opportunity during periods of slow sales to perform such extra duties, but treated the duties as voluntary in conformance with a norm that had developed under the previous management.

Noncompliant behaviors typically involved violating rules. For example, salespeople sometimes made unrealistic promises to customers in order to close a sale (e. g., fast delivery, inaccurate product characteristics, misstated company policies). This would frequently result in extra work or problems for other employees. Management administered sanctions according to the severity of the noncompliant behavior.

The prosocial items were factor analyzed separately from the noncompliant items to preserve the managers' conception of each type of behavior, inasmuch as the nature of prosocial and noncompliant behav-

iors can vary with the organizational context. The five prosocial items formed a single factor in a principal factor analysis with communalities on the diagonal of the correlation matrix, and were averaged. The five noncompliant behaviors also loaded on a single factor and were averaged (Table 1). The two scales were negatively correlated ($r = -.74, p < .001$).

Work performance. Sales figures for each respondent were obtained for the sales quarter ending with the month the survey was conducted. The 3-month period was selected to give a more stable representation of sales performance than that provided by a single month. Two adjustments were made to the gross sales figures. First, each person's sales were adjusted for the number of hours worked. Second, given the wide variance in sales volume, store size, and sales floor area across the 12 stores, z scores were calculated for each salesperson by subtracting from their sales the mean sales of the store and dividing by the standard deviation.

Need for achievement and need for autonomy. The two subscales from the Manifest Needs Questionnaire (MNQ; Steers & Braunstein, 1976) were selected because they were based on Murray's (1938) definitions of need for achievement and need for autonomy. Each scale was entered into a separate principal factor analysis after eliminating the two reverse-scored items due to low applicability to the sales profession. The four items representing need for achievement (e. g., "I try very hard to improve on my past performance at work") were averaged (eigenvalue = 1.92; 48% variance explained), as were the four items representing need for autonomy (e. g., "I go my own way at work regardless of the opinions of others"; eigenvalue = 1.95; 49% variance explained).

Satisfaction with material rewards. The two subscales, Satisfaction with Pay and Satisfaction with Security, were taken from Hackman and Oldham's (1975) instrument, Satisfaction with the Work Context. The four items (e. g., "The degree to which I am fairly paid for what I contribute to this organization," "How secure things look for me in the future in this organization") produced a single factor (eigenvalue = 2.76; 69% variance explained), and were averaged.

Perceived peer competition. Steers's (1975) two items measuring perceived peer competition (e. g., "There is a very competitive atmosphere among my peers and myself with regard to attaining our respective sales goals") were averaged.

Faith in peers. The three items composing Cook and Wall's (1980) scale of Faith in Peers (e. g., "I can trust the people I work with to lend me a hand if I need it") loaded on a single factor (eigenvalue = 2.41; 80% variance explained), and were averaged.

Confidence in management. Respondents rated their immediate supervisor on a seven-item scale that included five items from Cook and

Wall's (1980) Trust of Management scale and two items developed to assess perceptions of disciplinary style (i. e., "Management is prepared to take corrective action when employees violate rules and procedures," and "Management administers rewards and punishments fairly"). The seven items loaded on a single factor (eigenvalue = 2.41; 80% variance explained), and were averaged.

Results

Means, standard deviations, internal consistency estimates, and correlations of all variables are reported in Table 2.

The variables were entered simultaneously into regression analyses as reported in Table 3. Equations were statistically significant for both prosocial behavior, $F(6, 102) = 2.47, p < .05$, and for noncompliant behavior, $F(6, 123) = 2.39, p < .05$. Prosocial behavior was positively related to need for achievement and satisfaction with material rewards, and negatively related to perceived peer competition. Noncompliant behavior was negatively related to need for achievement and confidence in management.

Given the relatively high correlation between prosocial and noncompliant behavior, a canonical correlation analysis was performed. As shown in Table 4, the pattern was similar to the regressions, but the canonical variate for noncompliant behavior was not significant ($F = 1.65, p = .15$). Using .30 as a cutoff for significance, prosocial behavior had the highest loading on the first variate ($F = 1.95, p = .03$), along with satisfaction with material rewards, need for achievement, and perceived peer competition. Unlike in the regression, need for autonomy had a significant negative loading as hypothesized. Noncompliant behavior also was significant but only about half as large as prosocial. The second canonical variate showed noncompliant behavior to be negatively related to need for achievement and confidence in management. The regressions and canonical correlation analysis suggest that the achievement motive and personal security account for prosocial behavior, and low achievement and low reciprocity account for noncompliant behavior.

Zero-order correlations showed that work performance was positively related to prosocial behavior and negatively related to noncompliant behavior. In light of the relatively high correla-

Table 2

Means, Standard Deviations, Internal Consistency Estimates, and Correlations of All Variables

Variable	1	2	3	4	5	6	7	8	9
1. Need for achievement	(.60)								
2. Need for autonomy	.17**	(.65)							
3. Satisfaction with material rewards	.13*	-.23**	(.85)						
4. Perceived peer competition	.25**	.09	.20**	(.80)					
5. Faith in peers	-.02	-.26**	.34**	.20**	(.88)				
6. Confidence in management	-.03	-.29**	.35**	.40**	.35**	(.83)			
7. Prosocial behavior	.13*	-.18**	.27**	-.08	.10	.10	(.81)		
8. Noncompliant behavior	-.17**	.11*	-.16**	-.04	-.13*	-.23**	-.74**	(.79)	
9. Sales performance	.19**	.09	.23**	.17**	.14**	.15**	.16**	-.23**	(—)
<i>M</i>	5.41	4.07	4.97	5.59	5.47	5.03	4.80	2.66	.00
<i>SD</i>	0.90	1.12	1.34	1.35	1.24	1.34	1.07	0.92	.96

Note. $N = 141$. Coefficient alpha reliability coefficients are shown on the diagonal (for perceived peer competition, Spearman-Brown used correcting for $k = 2$). All variables are measured on a 7-point response scale except sales performance (z scores).

* $p < .10$. ** $p < .05$, one-tailed.

Table 3
Regression Results for Prosocial and Noncompliant Behaviors

Independent variable	Behavior	
	Prosocial	Noncompliant
Need for achievement	(+) .19*	(-) -.24*
Need for autonomy	(-) -.10	(+) .04
Satisfaction with material rewards	(+) .23*	(0) -.00
Perceived peer competition	(-) -.21*	(0) .09
Faith in peers	(+) .04	(-) -.01
Confidence in management	(+) .03	(-) -.26*
<i>F</i>	2.47*	2.39*
<i>R</i> ²	.13	.10
Adjusted <i>R</i> ²	.08	.06

Note. Entries are standardized regression coefficients. Hypothesized relationships are noted in parentheses.

* $p < .05$, one-tailed.

tion between prosocial and noncompliant behavior, a hierarchical regression analysis was performed. Two regressions were run with alternate ordering of entry of the two behaviors. The change in R^2 for each regression was then tested for significance. This procedure measured the independent effects of the two behaviors on performance by exploring the relation between one behavior and work performance with the variance shared by the other behavior and work performance partialled out. As shown in Table 5, noncompliant behavior had a greater incremental influence on performance than did prosocial behavior, which had no impact (.02 vs. .00 change in R^2). This suggests that noncompliant behavior undermines performance, but prosocial behavior neither helps nor hinders performance.

An additional regression was run to determine whether the independent variables were related to performance after the effect of prosocial and noncompliant behavior was partialled out. Prosocial and noncompliant behavior were entered jointly in the first step, followed by the six independent variables in the second step. As shown in Table 6, the change in R^2 at the second step was significant $F(8, 100) = 2.67, p < .05$. In addition to the significant negative effect of noncompliant behavior, need for achievement and satisfaction with material rewards were positively related to sales performance. To eliminate the inflated estimate of R^2 due to five nonsignificant variables, a regression containing only the three significant predictors was run. R^2 decreased from .19 to .11, and adjusted R^2 decreased from .12 to .09. These results show that need for achievement and satisfaction with material rewards are able to account for variations in performance over and above what is explained by noncompliant behavior.

Discussion

The findings suggest that prosocial and noncompliant behaviors are distinct types of nontask behavior that have a common achievement-motivation base but are influenced by different perceived situational contingencies.

Achievement-oriented individuals appear to actualize their need to excel by performing tasks outside their focal tasks that

Table 4
Canonical Correlation Analysis of Behaviors and Predictors

Variable	Canonical variates	
	1	2
Outcome		
Prosocial behavior	.96	-.28
Noncompliant behavior	-.52	.85
Predictor		
Need for achievement	.30	-.51
Need for autonomy	-.40	.12
Satisfaction with material rewards	.73	.04
Perceived peer competition	-.32	-.24
Faith in peers	.25	-.15
Confidence in management	.10	-.73
Canonical correlations	.36	.27
<i>F</i>	1.95*	1.65
% of variance	.65	.35
Eigenvalue	.15	.08

Note. Overall equation: $F = 1.95$.

* $p < .05$.

benefit the organization and by refraining from nonfocal activities dysfunctional to the organization. These results suggest that achievement-oriented individuals exhibit a consistent behavior pattern across different types of behavior.

The inconclusive results for need for autonomy may have been due to low reliability of the MNQ (see Blackburn, 1981, and Dreher & Mai-Dalton, 1983) or low applicability of the items to the sales profession.

The results fully supported hypotheses for personal security, confirming that personal security is a prerequisite to prosocial behavior but has no bearing on noncompliant behavior. The positive relation between prosocial behavior and satisfaction with material rewards paralleled that of altruism and job satisfaction in the C. Smith et al. (1983) study. The negative relation between prosocial behavior and perceived peer competition was consistent with Tjosvold's (1986) model of cooperative and competitive goal interdependence. The results support Duval and Wicklund's (1972) self-awareness theory, which argues that people are less preoccupied with themselves when they feel personally secure. The nonsignificant findings for noncompliant

Table 5
Hierarchical Regressions for Work Performance

Independent variable	<i>R</i> ²	Adjusted <i>R</i> ²	<i>F</i>	<i>R</i> ² change	<i>F</i>
Regression 1					
Step 1. Prosocial behavior	.03	.02	2.98*	—	—
Step 2. Noncompliant behavior	.05	.04	3.15**	.02	3.27*
Regression 2					
Step 1. Noncompliant behavior	.05	.05	6.29**	—	—
Step 2. Prosocial behavior	.05	.04	3.15**	.00	.07

* $p < .10$. ** $p < .05$.

Table 6
Regression of Work Performance on All Variables

Independent variable	Stan- dardized β coef- ficients	R^2	Adjusted R^2	F	R^2 change	F
Step 1.						
Prosocial behavior	-.03					
Noncompliant behavior	-.26**					
		.06	.04	3.03**	—	—
Step 2.						
Need for achievement	.15*					
Need for autonomy	.12					
Satisfaction with material rewards	.22**					
Perceived peer competition	.06					
Faith in peers	.00					
Confidence in management	.11					
		.19	.12	2.83**	.13	2.67**

* $p < .10$. ** $p < .05$, one-tailed.

behavior and personal security are consistent with the hypothesis that people refrain from noncompliant behavior out of fear of further jeopardizing their personal security by being sanctioned.

Support for the reciprocity explanation was obtained for noncompliant behavior but not for prosocial behavior. Although surprising from a conceptual perspective, the nonsignificant effect for prosocial behavior paralleled C. Smith et al.'s (1983) finding of no direct link between leader supportiveness and altruism. Had the prosocial behaviors included more that benefited specific individuals directly rather than the organization in general, perhaps a significant relation might have emerged (Brief & Motowidlo, 1986). The negative relation between confidence in management and noncompliant behavior was consistent with C. Smith et al.'s (1983) positive relation between leader supportiveness and generalized compliance. These results suggest that violation of management's rules is a way of reciprocating poor treatment by one's supervisor.

The relation of prosocial and noncompliant behavior to work performance must be interpreted with caution. The negative relation between noncompliant behavior and work performance supports the hypothesis that work rules are imposed with the aim of encouraging high performance. The modest relation between prosocial behavior and work performance suggests that prosocial behaviors, although considered desirable by management, provided intangible benefits to the organization (e. g., enhancement of the company's reputation and goodwill, or positive affect among employees) rather than monetary benefits.

However, alternate explanations may account for the results. Rater bias due to the halo effect might apply in that supervisors may have allowed their knowledge of an individual's sales per-

formance to color their ratings of prosocial and noncompliant behavior. An additional ambiguity is the direction of causality. Not only was the study cross sectional, but the performance variable was a composite of sales figures for the month of the study as well as for the 2 preceding months. Although this questions sales performance as a dependent variable, the correlation ($r = .68, p < .0001$) of this quarterly performance measure with performance over the preceding 12-month period suggests that performance is relatively stable and would be similar in future.

The regression of work performance on all variables has several implications. The positive relation with need for achievement confirms the well-established finding that the reward and feedback contingencies of the sales profession are well suited to people high in need for achievement (e. g., Atkinson, 1964; see Churchill, Ford, Hartley, & Walker, 1985, for a meta-analytic review). The positive relation with satisfaction with material rewards suggests that personal security enhances performance by reducing preoccupation with the self; however, an equally plausible explanation is simply that high performance increases material rewards. The fact that the remaining four variables were unrelated to work performance suggests that other variables, such as task behaviors, may be important moderators.

I have argued that prosocial behavior and noncompliant behavior are separate constructs. The alpha reliability coefficients provide evidence of internal consistency. However, the relatively strong negative correlation of prosocial and noncompliant behavior questions whether they are simply opposite poles of a single construct, ranging from a low in which people do things that are wrong to a high where they go above and beyond the call of duty. Some support for discriminant validity comes from the results for the two personal security variables, whereby as predicted, significant relations were found with prosocial behavior and null relations were found for noncompliant behavior. It would be useful to test other variables suspected of following this pattern. Comparing antecedents of focal task behaviors and nontask behaviors would further clarify the discriminant validity issue. Nevertheless, my findings suggest a dual-factor theory (Herzberg, 1966) of nontask behaviors that proposes that the opposite of prosocial behavior is no prosocial behavior, rather than noncompliant behavior, and that the opposite of noncompliant behavior is compliant behavior, not prosocial behavior. The fact that some variables relate to prosocial and noncompliant behavior in opposite directions does not necessarily preclude prosocial and noncompliant behavior from being separate constructs in that the process underlying the relations or the moderating variables may differ.

The findings, although modest, suggest that managers should be aware that noncompliant behavior can have important consequences for work performance, as can need for achievement and satisfaction with material rewards. In addition, managers should not assume that conditions discouraging noncompliant behavior do not necessarily promote prosocial behavior.

References

Atkinson, J. W. (1964). *An introduction to motivation*. Princeton, NJ: Van Nostrand.
Bateman, T. S., & Organ, D. W. (1983). Job satisfaction and the good soldier: The relationship between affect and employee 'citizenship.' *Academy of Management Journal*, 26, 587-595.

- Berkowitz, L., & Daniels, L. R. (1963). Responsibility and dependency. *Journal of Abnormal and Social Psychology*, 66, 429-437.
- Berkowitz, L., & Daniels, L. R. (1964). Affecting the salience of the social responsibility norm: Effects of past help on the response to dependency relationships. *Journal of Abnormal and Social Psychology*, 68, 275-281.
- Beyer, J. M., & Trice, H. M. (1984). A field study of the use and perceived effects of discipline in controlling work performance. *Academy of Management Journal*, 27, 743-764.
- Blackburn, R. (1981). An evaluation of the reliability, stability, and factor structure of the M.N.Q. *Journal of Management*, 7(2), 55-62.
- Brief, A. P., & Motowidlo, S. J. (1986). Prosocial organizational behaviors. *Academy of Management Review*, 11, 710-725.
- Bunker, D. R., Banckert, L., Di Biase, K., & McGillicuddy, N. (1986). *Promoting the persistence of organizational transformations*. Unpublished manuscript, State University of New York at Buffalo, School of Management.
- Burke, R. J. (1982a). Personality, self-image, and situational characteristics of effective helpers in work settings. *The Journal of Psychology*, 112, 213-220.
- Burke, R. J. (1982b). Personality, self-image, and informal helping processes in work settings. *Psychological Reports*, 50, 1295-1302.
- Burke, R. J., & Weir, T. (1978). Organizational climate and informal helping processes in work settings. *Journal of Management*, 2, 91-105.
- Burke, R. J., Weir, T., & Duncan, G. (1976). Informal helping relationships in work organizations. *Academy of Management Journal*, 19, 371-377.
- Churchill, G. A., Ford, N. M., Hartley, S. W., & Walker, O. C. Jr. (1985). The determinants of salesperson performance: A meta-analysis. *Journal of Marketing Research*, 22, 103-118.
- Cook, J., & Wall, T. D. (1980). New work attitude measures of trust, organizational commitment and personal need nonfulfillment. *Journal of Occupational Psychology*, 53, 39-52.
- Dovidio, J. F. (1984). Helping behavior and altruism. An empirical and conceptual overview. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 361-427). New York: Academic Press.
- Dreher, G. F., & Mai-Dalton, R. R. (1983). A note on the internal consistency of the Manifest Needs Questionnaire. *Journal of Applied Psychology*, 68, 194-196.
- Duval, S., & Wicklund, R. A. (1972). *A theory of objective self-awareness*. New York: Academic Press.
- Eisenberger, R., Huntington, R., Hutchison, S., & Sowa, D. (1986). Perceived organizational support. *Journal of Applied Psychology*, 71, 500-507.
- Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review*, 25, 161-178.
- Hackman, J. R., & Oldham, G. R. (1975). Development of the Job Diagnostic Survey. *Journal of Applied Psychology*, 60, 159-170.
- Herzberg, F. (1966). *Work and the nature of man*. Cleveland, OH: World Publishing.
- Iaffaldano, M. T., & Muchinsky, P. M. (1985). Job satisfaction and job performance: A meta-analysis. *Psychological Bulletin*, 97, 251-273.
- Kaplan, E. M., & Cowen, E. L. (1981). Interpersonal helping behavior of industrial foremen. *Journal of Applied Psychology*, 66, 633-638.
- Katz, D. (1964). The motivational basis of organizational behavior. *Behavioral Science*, 3, 131-146.
- Miller, D. T. (1977). Personal deserving versus justice for others: An exploration of the justice move. *Journal of Experimental Social Psychology*, 13, 1-13.
- Motowidlo, S. J. (1984). Does job satisfaction lead to consideration and personal sensitivity? *Academy of Management Journal*, 27, 910-915.
- Murray, H. A. (1938). *Explorations in personality*. New York: Oxford University Press.
- O'Reilly, C. A. III, & Chatman, J. (1986). Organizational commitment and psychological attachment: The effects of compliance, identification, and internalization on prosocial behavior. *Journal of Applied Psychology*, 3, 492-499.
- O'Reilly, C. A. III, & Weitz, B. A. (1980). Managing marginal employees: The use of warnings and dismissals. *Administrative Science Quarterly*, 25, 467-484.
- Podsakoff, P. M., & Todor, W. D. (1985). Relationships between leader reward and punishment behavior and group processes and productivity. *Journal of Management*, 11, 55-73.
- Podsakoff, P. M., Todor, W. D., & Skov, R. B. (1982). Effects of leader contingent and noncontingent reward and punishment behaviors on subordinate performance and satisfaction. *Academy of Management Journal*, 25, 810-821.
- Rosenhan, D. L., Salovey, P., & Hargis, K. (1981). The joys of helping: Focus of attention mediates the impact of positive affect on altruism. *Journal of Personality and Social Psychology*, 40, 899-905.
- Schneider, B. (1985). Organizational behavior. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual review of psychology* (pp. 573-611). Stanford, CA: Annual Reviews.
- Smith, C. A., Organ, D. W., & Near, J. P. (1983). Organizational citizenship behavior: Its nature and antecedents. *Journal of Applied Psychology*, 68, 653-663.
- Smith, F. J. (1977). Work attitudes as predictors of attendance on a specific day. *Journal of Applied Psychology*, 62, 16-19.
- Steers, R. M. (1975). Task-goal attributes, achievement, and supervisory performance. *Organizational Behavior and Human Performance*, 13, 392-403.
- Steers, R. M., & Braunstein, D. N. (1976). A behaviorally-based measure of manifest needs in work settings. *Journal of Vocational Behavior*, 9, 251-266.
- Tjosvold, D. (1986). *Working together to get things done: Managing for organizational productivity*. Lexington, MA: Heath.
- Weiss, H. M. (1984). Contributions of social psychology to productivity. In A. P. Brief (Ed.), *Productivity research in the behavioral and social sciences*. (pp. 143-173). New York: Praeger.
- Wiener, Y., & Gechman, A. S. (1977). Commitment: A behavioral approach to job involvement. *Journal of Vocational Behavior*, 10, 47-52.
- Williams, L. J., Podsakoff, P. M., & Huber, V. (1986, August). *Determinants of organizational citizenship behaviors: A structural equation analysis with crossvalidation*. Paper presented at the meeting of the Academy of Management, Chicago.
- Worthy, J. C. (1986, August). *Overachievement at work: A class of prosocial behavior*. Paper presented at the meeting of the Academy of Management, Chicago.

Received November 10, 1986

Revision received February 12, 1987

Accepted April 2, 1987 ■

Business Climate Attitudes and Company Relocation Decisions

Neal Schmitt, Sandra E. Gleason, Bruce Pigozzi, and Philip M. Marcus
Center for Redevelopment of Industrial States, Michigan State University

A total of 438 manufacturing firms reported their business relocation activity, their overall attitude toward the business climate in Michigan, and their assessment of 34 business climate dimensions. Using a lens model paradigm, we explored reasons for a lack of relation between actual relocation decisions and overall business climate attitudes. Location decisions were most highly correlated with distance and labor considerations, whereas overall business climate was most highly correlated with tax considerations, and secondarily correlated with labor problems.

Studies of the attitudes toward the business climate in various states and regions of the country have become increasingly popular in recent years. Because of the perceived high economic stakes involved, these studies frequently receive considerable political and media attention. The product of most of these studies is a rank order of states or localities in terms of their favorable treatment of business (Mattila & Thompson, 1983).

Perhaps the most popular—as well as the most frequently criticized—study of business climate is that produced annually since 1979 by the Alexander Grant Company. Factors to be considered and the factor weights are provided by members of the Conference of State Manufacturers' Associations. Although the weights change from year to year, some variety of state and local government fiscal policies, state-regulated employment costs, labor costs, availability and productivity of the labor force, and other manufacturing-related costs thought to be important for business success, such as energy and environmental controls, are included. The values for each of these factors for each of the 50 states are extracted from published sources such as the U.S. Census of Population, normalized, multiplied by the agreed upon weights, and then summed. The summed products are used as a basis to rank states by their business climate "score."

Aside from the expected political and emotional responses of officials in low- and high-ranked states, these state business climate studies have generated a variety of scientific or measurement criticisms. First, the cost factors included have been criticized as being of primary importance only to mass production industries and firms using low-skilled workers available in less industrialized areas. Furthermore, it is argued that the factors overrepresent items related to labor and exclude factors important to businesses other than manufacturing production firms, such as small entrepreneurs, corporate headquarters facilities, and research and development firms. By contrast, Thompson and Thompson (1983) have attempted to use such indices as

the percentage of minorities employed in administrative and executive positions, the number of airplane departures, and the number of earned doctorates in local universities to derive a business climate index more appropriate for nonmanufacturing companies. Another criticism relates to the fact that many of the aggregate state statistics used show great variability within states across industry and locality (see Hunt, 1985, for an excellent example). Finally, previous studies have necessarily been limited to assessments of factors for which published data exist, such as economic or labor force statistics. As a consequence, these studies often ignored less tangible factors such as quality of life factors.

To these criticisms, we might add that no previous study has examined the reliability and intercorrelation of the importance indices used to weight various factors. No previous study has pursued the question of whether these indices, or other attitudinal measures of a state's business climate, are related to any actual or intended behavior on the part of organizations. The latter point is particularly important inasmuch as a considerable body of literature indicates that individuals' attitude and behavior are frequently uncorrelated (e.g., Fishbein & Ajzen, 1975). Finally, most, if not all, previous research has been atheoretical.

This article addresses some of these issues through a study of the attitudes and behaviors of Michigan manufacturing businesses. Specifically, we attempt to relate the assessment of business executives regarding their present location in Michigan to their actual decision to build new facilities in or out of Michigan and their overall impression of the business climate in Michigan.

In the current article, we report the development of measures of various dimensions of business climate. We then use a policy-capturing method and a lens model paradigm (Slovic & Lichtenstein, 1971) to investigate four issues: (a) the degree to which impressions of individual business climate dimensions relate to overall assessments of business climate, (b) the degree to which perceptions of climate dimensions relate to actual relocation decisions, (c) the degree to which policies derived from subjects' responses using actual decisions and perceived overall climate are similar, and (d) the degree to which subjective weights for various business climate dimensions match either of the statistically derived weights or whether they are more or less predictive of actual behavior than are the statistically derived weights.

The authors would like to thank Michael Doherty and an anonymous reviewer for their helpful comments on two earlier versions of this article.

Correspondence concerning this article should be addressed to Neal Schmitt, Department of Psychology, Michigan State University, East Lansing, Michigan 48824-1117.

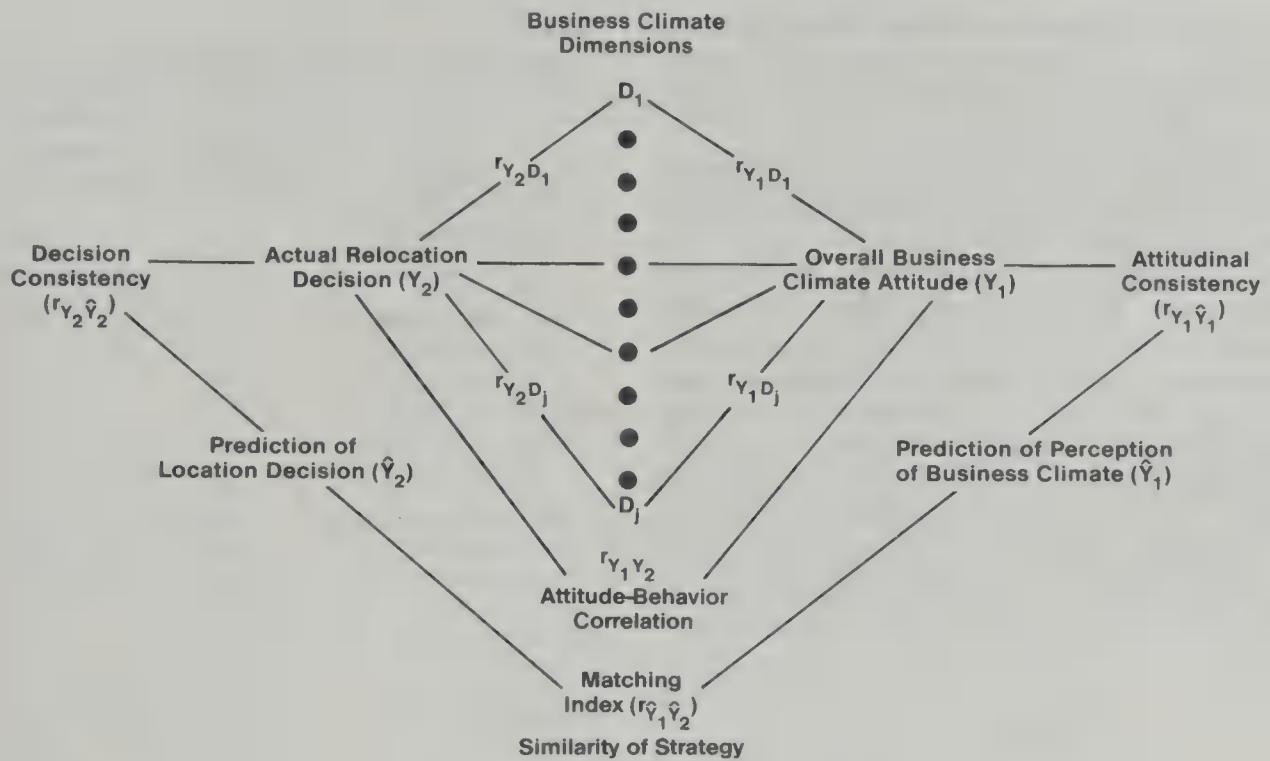


Figure 1. Lens model representation of business climate attitudes and relocation decision.

The lens model paradigm that guided the research and provided the indices described in this article is depicted in Figure 1. The model makes the assumption that business decision makers have attitudes or information about several different dimensions of the business climate such as the availability and quality of labor, corporate taxes, and the quality of life. These individual impressions are related to their overall impression of the business climate. A regression of the overall business climate on the individual climate dimensions yields statistical weights (Schmitt & Levine, 1977; Stumpf & London, 1981) for each of the climate dimensions. In the absence of "high" intercorrelations among the climate dimensions and given that there are not major missing dimensions, these beta weights can be viewed as indices of the importance of individual climate dimensions in a determination of overall business climate. The correlation ($r_{Y_1Y_2}$) is an index of the degree to which the overall assessment of climate is consistent with the weighted composite of their evaluation of individual dimensions. Similarly, a regression of the actual location of business facilities on assessments of individual climate dimensions yields statistical weights of the importance of these climate dimensions in actual relocation decisions. The correlation ($r_{Y_2\hat{Y}_2}$) represents the consistency between actual location decisions and the statistically weighted composite of respondents' assessments of individual climate dimensions. The correlation ($r_{Y_1Y_2}$) between overall business climate attitudes and the actual location decision is an index of the "validity" of the overall attitude measure. More broadly, this is an attitude-behavior correlation; it represents the validity of the overall evaluation of business climate as a predictor of location decisions.

The matching index ($r_{\hat{Y}_1\hat{Y}_2}$) depicted in the diagram represents the correlation between predicted values obtained with the two regression equations. If this correlation is relatively high, it indicates a similarity between the two sets of statistically

derived weights. The attitude-behavior correlation ($r_{Y_1Y_2}$) should be equal to the product of the two consistency indices ($r_{Y_1\hat{Y}_1}$ and $r_{Y_2\hat{Y}_2}$) and the matching index ($r_{\hat{Y}_1\hat{Y}_2}$), assuming a linear additive model (see Slovic & Lichtenstein, 1971; or more recently, Schmitt, Noe, & Gottschalk, 1986) and that all relevant predictors have been measured. Examination of these various indices not only allows a determination of the degree to which attitude and behavior are consistent, but also allows some determination of how and why they are inconsistent. If consistency indices are low, it means attitudes regarding the climate dimensions *measured* are not major determinants of either overall attitude or actual location behavior. Or, if consistency indices are high and matching is low, one would conclude that decision makers are using the information they have on individual climate dimensions to make location decisions and to form overall assessments of climate, but that they are using the information very differently in the two cases. If matching is low, but consistency is high, it would suggest that although business decision makers' unhappiness with certain aspects of the business climate is reflected in their overall attitude concerning the climate, it is either impossible for them to act consistently with these feelings or they are unwilling to do so.

It is important to note that the procedure and indices previously described represent a modification of the usual lens model study. In a typical lens model study, each individual is asked to respond to a series of cases, each case being represented by a unique set of D_i . Their responses to these D_i are used to compute the various indices described; hence, the individual judge is the unit of analysis. In the study described in this article, a group of judges (business executives) each provides a judgment of a single case based on their assessments of the unique D_i in their situation. They also indicate their attitude regarding the overall business climate that serves as one dependent variable (Y_1) and the information regarding business relocations

(Y₂), which is the second dependent variable. The judge in the procedure outlined is actually a group of judges. Heterogeneity among the individuals' decision-making processes will obviously lower the various lens model indices.

The fourth issue addressed in this article is whether subjective weights of the importance of the business climate dimensions coincided with either set of statistically derived weights. We computed the parallel of a consistency index for subjective weights; that is, the correlation between the subjectively weighted sum of the climate dimensions and the overall assessment of business climate and actual business location decisions. Furthermore, this subjectively weighted composite was correlated with predicted values (\hat{Y}_1 and \hat{Y}_2) from the two regression equations to yield an index of the comparability of these various sets of weights. This index is directly parallel to the matching index already described. Previous comparisons of objective and subjective weights indicate only moderate agreement (Cook & Stewart, 1975; Slovic & Lichtenstein, 1971; Zedeck & Kafry, 1977).

Method

Sample

Surveys were mailed to 950 manufacturing members of the Michigan Chamber of Commerce. Nonrespondents were sent a follow-up survey 3 weeks after the first survey was mailed. In all, 438 usable surveys (46%) were obtained. The addressee was instructed to answer the questionnaire if he or she had had experience with plant relocation or expansion decisions; if not, the instrument was to be forwarded to someone in the firm who could provide informed responses. Of the respondents, 48% were the president or chief executive officer of their firm, 18% were vice-presidents, 10% were plant managers, 10% were treasurers, 9% were other executive officers in the organization, and the remainder held a variety of other offices.

The addressee sample comprised all Chamber of Commerce members in six manufacturing Standard Industrial Classification (SIC) categories as indicated in Table 1. Table I includes data on the population of organizations in Michigan, the organizations to whom surveys were sent, and the respondent organizations. The Chamber membership was approximately the same as that of the state of Michigan in regard to the proportion of firms engaged in the six types of manufacturing, but the respondents overrepresented fabricated metals firms and underrepresented machinery manufacturers. Large firms were overrepresented in both the respondent sample and the Chamber of Commerce membership relative to the U.S. Department of Commerce (1982) census figures. However, the census data included many self-employed persons as businesses, and these respondents were of no interest in our study. Finally, the analyses reported ahead included only those who reported some actual or planned expansion or relocation.

Measures

Dependent Variables. Our main interest was in determining whether and how the perceptions of the state's business climate relative to other states were related to actual decisions about relocation and overall business climate attitudes in Michigan. The dependent variables were provided by answers to two questions. Actual relocation decisions were measured by asking the respondents whether their company had opened any new facilities in the past 5 years. If they answered *yes*, they were then asked to indicate the state in which the activity took place. If the expansion took place in Michigan, the variable was coded 2; if elsewhere it was coded 1. The potential sample size included the 187 companies

Table 1
Sample Characteristics (in Percentages)

Characteristic	Organization population ^a	Chamber of Commerce membership ^b	Respondent sample
Standard Industrial Classification category ^c			
Food and kindred (20)	5.4	8.0	7.9
Lumber, wood, furniture, paper (24, 25, 26)	11.9	12.0	12.3
Chemicals, rubber, plastics etc. (28, 29, 30, 32, 33)	19.7	27.0	19.4
Fabricated metals (34)	20.2	20.0	42.9
Machinery (35, 36)	38.2	27.0	11.9
Transportation (37)	4.5	8.0	5.5
Firm size (no. of employees)			
0-100	91.0	40.4	37.2
100-500	7.4	49.4	48.4
500-1000	0.8	7.2	7.6
1000 +	0.8	3.0	6.7
Sample size	11,176	950	438

^a U.S. Department of Commerce. (1984). Table 1B.
^b Michigan Chamber of Commerce, 1984 membership list.
^c Numbers in parentheses represent the Standard Industrial Classification code.

who reported some expansion activity. A total of 92 of these 187 firms reported building in Michigan, and 95 elsewhere. Note that few companies actually closed plants in Michigan to move to another location (relocation), and we had no way to differentiate these companies from those that maintained facilities in Michigan but opened new facilities either in Michigan or elsewhere. The second dependent variable was provided by answers to the question "How would you compare the overall business climate of Michigan to that of other states? Michigan's business climate is . . ." The 5-point response scale ranged from *a great deal better* (1) to *a great deal worse* (5). The mean and standard deviation of this measure were 4.39 and .73, respectively.

Subjective weights. Respondents were also asked to indicate their view of the importance of 34 business climate items (see the list in Table 2). These items were derived from the previous research on business climate, as well as from interviews with company personnel, Michigan Chamber of Commerce employees, Michigan Department of Commerce employees, and utility company representatives. The 34 items represented a reasonably comprehensive list of potential factors that firms consider when making decisions regarding relocation or expansion. It is certainly true that other more detailed information about a location might be sought, such as was represented by the Thompson and Thompson (1983) study cited earlier, but it was felt that the 34 items used were generally applicable and were those most frequently used in previous studies.

Company representatives were asked to indicate the importance of the factors in business location decisions on a 6-point scale ranging from *the most important factor* to *a factor of no importance at all* (Importance). These responses were used as subjective importance weights for the various climate dimensions.

Predictor variables. Respondents were also asked to compare their location within the state of Michigan to a location outside of the state in which they had recently located a facility *or* to the most attractive outside-the-state location (Comparison). This comparison was made on

Table 2
Scale Items, Means, and Standard Deviations
of Business Climate Dimensions

Item	M_C	SD_C	M_I	SD_I
Taxes	1.90	0.95	3.98	1.08
State taxes on business				
Local taxes on business				
State/local taxes on individuals				
Costs of workers compensation				
Costs of unemployment compensation				
Natural Resources	3.18	0.83	3.24	0.96
Ample area for future expansion				
Costs of property and construction				
Water supply and costs				
Availability and cost of energy				
Zoning and other regulations				
Environmental protection requirements				
Labor	2.58	0.98	4.13	0.94
Availability of unskilled or semiskilled				
Productivity of workers				
Wage rates				
Labor relations				
Extent of worker unionization				
Quality of Life	3.38	0.77	2.83	0.89
Size of city or town				
Fiscal health of state				
Style of living for employees				
Cost of living				
Crime rate				
Personal preference of CEO				
Skilled Workers	4.45	1.28	3.37	1.26
Availability of skilled workers				
Availability of tech-prof workers				
Finance	3.01	1.07	3.09	1.40
Local sources of financing				
State-local financial inducements				
Distance	4.32	1.09	3.33	0.86
Distance to customers				
Distance to materials				
Distance to services				
Distance to other company facilities				
Transportation	4.07	0.86	2.82	0.98
Transportation facilities for people				
Transportation facilities for materials and products				

Note. M_C and SD_C refer to the scale means and standard deviations for the comparison ratings. M_I and SD_I refer to the importance means and standard deviations. CEO = chief executive officer.

a 7-point scale ranging from *Michigan location is very much better* to *out-of-state location is very much better*. Respondents were also allowed to indicate *do not know* in making their comparison ratings (this response was treated as missing data), and they were asked to indicate what alternative location they considered when making the ratings. These comparison ratings were taken as the D_i (see Figure 1) as they represented the respondents' assessment of Michigan's status on these dimensions. They were the predictor variables in the regression analyses that produced the consistency indices in Figure 1.

Analysis of dimensionality of climate variables. As indicated previously, one criticism of past business climate studies was the redundancy of the items used and the lack of attention to the measurement properties of the scales used. This was also a concern for our study inasmuch as several of the climate items were conceptually similar. For ex-

ample, wage rates and labor availability are both labor-related factors. Reducing this redundancy is desirable to avoid overrepresenting some factors (Thompson & Thompson, 1983) and permits meaningful multivariate analyses of the type to be described ahead.

To reduce this item redundancy, we used a combination of rational and empirical considerations (Nunnally, 1978), which produced eight business climate scales. First, content considerations were used as the basis for grouping the variables. Two of the original 34 items were not included in these eight scales. Business climate, attitudes toward industry was excluded because it was too similar to the overall business climate dependent variable. A second item (marketing facilities) did not correlate well with any of the eight clusters of items, nor did it correlate significantly with either dependent variable. It was also dropped from further analyses. In Table 3, we present the internal consistencies and intercorrelations of the eight dimensions for both the Comparison and Importance ratings. Although clearly not independent, intercorrelations of dimensions are much lower than internal consistencies providing empirical support for our content clusters.

A score for each of the eight scales based on these empirical and rational considerations (see Table 2 for a description of the items in each scale) was computed as the arithmetic average of the summed ratings for which an individual provided responses.

Data Analysis

Regression analyses of the actual location decision and the overall attitude regarding the Michigan business climate on the eight business climate dimensions provided measures of decision consistency and attitudinal consistency, as well as objective indices of the importance of each of the eight dimensions (see Figure 1). The two regression equations were used to compute predicted values (\hat{Y}_1 and \hat{Y}_2) for the overall climate attitude and the location decision, respectively. Subjective importance weights (the summed values of the Importance ratings for each scale in Table 2) were multiplied by the Comparison ratings and added to produce an index by which to compare the subjective weights and the two sets of objective weights. Pearson product-moment correlations were used to provide matching indices and a measure of the attitude-behavior relation.

Results

In Table 4, we present the results of the two multiple regression analyses. Both multiple correlations were statistically significant ($p < .05$), but of moderate size. A comparison of the beta weights and zero-order correlations indicates that there was little problem with multicollinearity. Comparison of beta weights across the two sets of analyses reveals little similarity. The actual location decision seems to be driven primarily by labor and distance considerations, whereas the overall assessment of the business climate index is primarily a function of taxes and, to a lesser extent, labor.

The results of analyses on the subjective importance ratings are presented in Table 5. Because of scale differences, the subjective importance weights and the two sets of "objective" weights were rank ordered to provide a comparison of subjective and objective weights. As can be seen, there are important differences in the rank order of these three sets of weights. These rank orders provide only rough comparisons, whereas the intercorrelations of predicted values from these correlations (the matching index) yield a single index of the similarity among the three weighting strategies. These intercorrelations, as well as correlations with the actual location decision and the overall

Table 3
Scale Reliabilities, and Scale Intercorrelations

Factor	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Importance																
1. Tax	.90															
2. Nature	.39	.83														
3. Labor	.41	.29	.77													
4. Life	.52	.44	.32	.78												
5. Skilled workers	.18	.11	.27	.27	.79											
6. Finance	.43	.43	.26	.50	.15	.77										
7. Distance	-.03	.16	.18	.15	.15	.10	.50									
8. Transportation	.18	.34	.28	.35	.24	.23	.24	.61								
Comparison																
9. Tax	-.17	-.02	-.12	-.02	-.02	-.06	.10	.06	.91							
10. Nature	.04	-.02	.01	.11	.09	.03	.09	.14	.41	.73						
11. Labor	-.09	.15	-.16	.12	.18	.09	.21	.17	.42	.36	.79					
12. Life	.02	.01	.00	.03	.08	.10	.11	-.03	.24	.46	.32	.60				
13. Skilled workers	.06	.00	.11	.11	.21	.07	.18	.08	.01	.14	.14	.20	.79			
14. Finance	.02	.03	-.07	.08	.12	-.20	.13	.06	.42	.46	.32	.37	.19	.64		
15. Distance	-.06	-.01	.06	.07	.20	.14	.26	.12	-.11	.04	.00	.05	.36	.12	.62	
16. Transportation	.10	.10	.18	.05	.07	.12	.10	.09	-.05	.10	.03	.21	.37	.14	.47	.73

Note. Internal consistency reliabilities are located on the diagonal.

business climate measure, are presented in Table 6. The attitude-behavior relation (the correlation between the location decision and the overall climate perception) is low (.12) and non-significant. This is not surprising given the relatively low level of consistency (.39 and .48) with which the eight climate dimensions were apparently used to come to overall evaluations of climate or a location decision. Moreover, the matching indices indicated a lack of correspondence between weights derived from different sources. The two sets of objective weights produced predicted values that were correlated only .38. The sum of the Importance-Comparison products (subjectively weighted climate dimensions) were correlated to some degree with both sets of predicted values generated from the regression equations. In this context, it is interesting to note that the corre-

lation (.48) of the subjectively weighted composite with predicted values coming from the actual decision regression (.44) was similar to the correlation (.46) of predicted values coming from the overall attitude equation and the same subjectively weighted composite.

Discussion

The results reported earlier indicate only a modest relation between evaluations of individual business climate dimensions and overall attitudes regarding business climate and actual business location decisions. There are several potential reasons why this relation was not stronger. First, the 32 items we used to assess business climate may not reflect all those factors that a business person(s) uses to decide on a particular location. However, an effort was made to be all inclusive in listing factors with-

Table 4
Results of Regression Analyses on Actual Location Decision and Perception of Overall Business Climate

Factor	Location decision		Overall climate	
	β	r	β	r
Taxes	-.06	-.04	.35*	.39*
Natural resources	-.12	.00	-.12	.09
Labor	.25*	.21*	.28*	.33*
Quality of life	.10	.12	.07	.16*
Skilled workers	-.12	.03	-.05	.02
Finances	.07	.10	.06	.17*
Distance	.30*	.27*	.04	-.00
Transportation	-.03	.09	.09	.04
R^2	.183*		.24*	
n^a	144		139	

^a n was less than 187 because of missing data on one or more of the variables.
* $p < .05$.

Table 5
Subjective and Objective Weights of Business Climate Dimensions

Dimension	Subj Importance M	Rank order		
		Subj	Obj _{AD} ^a	Obj _{JOP} ^a
Taxes	3.98	2	6	1
Natural resources	3.24	5	8	3
Labor	4.13	1	2	2
Quality of life	2.83	7	3	5
Skilled workers	3.37	3	7	7
Finances	3.09	6	4	6
Distance	3.33	4	1	8
Transportation	2.82	8	5	4

Note. Subj = subjective; Obj = objective.
^a Obj_{AD} indicates the rank order of the beta weights from the actual decision regression; Obj_{JOP} indicates the beta weights for the overall climate attitude regression.

Table 6

Intercorrelations Between Actual and Predicted Values of the Location Decision and Attitudes Regarding Business Climate

Variable	1	2	3	4	5
1. Location decision	—				
2. Overall climate attitude	.12 ^a	—			
3. Predicted decision	.39 ^b	.19*	—		
4. Predicted climate attitude	.14*	.48 ^b	.38 ^c	—	
5. Subjectively weighted composite	.22*	.29*	.44 ^c	.46 ^c	—

^a Indicates an attitude-behavior relation. ^b Indicates a consistency index. ^c Indicates a matching index.* $p < .05$.

out becoming very idiosyncratic to a particular company or group of companies' wishes. That we were relatively complete was supported by the fact that very few respondents listed any additional considerations when asked to do so in an open-ended question.

Second, the dichotomous nature of the business location dependent variable and the relatively low variance in the attitudinal variable may have served to lower the computed correlations.

Third, it is likely that different types of industry have very specific types of needs and wants, and our aggregation of responses from different companies served to hide those relations present. Some evidence that this was the case was revealed by comparing companies by SIC code (see Table 1). Discriminant analyses indicated that importance of natural resources and the availability of skilled personnel differed across industry type (Schmitt, 1986). Although this would certainly be an important distinction to make in subsequent research, it did not seem reasonable to subgroup the sample given the sample size available for regression analysis. This use of a "group" equation also represents the major modification in our use of the lens model relative to earlier studies of decision making. That is, in previous studies the individual made multiple judgments given different values of the D_i . In this study, multiple individuals representing different firms gave judgments about their situation. Their individual judgmental idiosyncrasies certainly served to lower the observed lens model indices in this study.

Finally, it is possible that respondents combine data in some nonlinear, configural manner to come to overall judgments or decisions, but the overwhelming research evidence indicates that nonlinear models of the decision process allow predictions that are no better than linear ones (Dawes & Corrigan, 1974). Process tracing studies that are more sensitive to nonlinear strategies of information processing may lead to better predictive models (e.g., see Billings & Marcus, 1983; Einhorn, Kleinmuntz, & Kleinmuntz, 1979).

The analyses did provide relatively definite answers to two of the questions that we proposed. First, the relation between overall attitudes toward the business climate and actual relocation behavior was near zero. This is, of course, consistent with much earlier literature on attitude-behavior relations. What is novel about the data presented here is that we have been able to indicate some reasons why this relation is low. Given linear assumptions and a complete list of predictor variables, the attitude-behavior correlation should be a product of the two consistency indices and the matching index (Schmitt et al., 1986; Slovic &

Lichtenstein, 1971). Both consistency indices were significant but modest in size. Furthermore, the matching index for the two regression equations was only .38, indicating that the two equations were not very similar. In making actual location decisions, companies appear to be most heavily influenced by distance considerations, and secondarily influenced by labor. Their overall attitude about the state business climate was most highly related to concerns about taxes and labor problems. What appears to be happening is that firms are complaining about the cost and perceived problems of doing business in the state, but because they need to be close to materials, customers, and services, they remain in Michigan. This interpretation seems appropriate given the dominance of the auto industry in Michigan and given the fact that many of the firms in our sample (note the SIC codes in Table 1) are likely auto suppliers. If these companies had alternatives, it is at least likely they would locate elsewhere. We may be observing the same phenomenon in this company relocation decision as organizational researchers have found in individual turnover research. That is, employees (or firms) may want to leave, but the economic situation precludes such action (Mobley, Horner, & Hollingsworth, 1978).

It is also important to note that even a correlation of .12 (the attitude-behavior relation in this article) may have important practical consequences. The utility of attitude-behavior relations, like that of ability-performance relations is directly proportional to the correlation (Brogden, 1946). In the case of an organization's decision to locate or expand in a given community, the economic circumstances can be enormous. Hence, even an r of .12 could have significant practical consequence.

There may also be states or geographic regions that are less dominated by the auto industry than is Michigan in which relocation decisions might be affected in different ways. We suggested earlier that some firms may be staying in Michigan because of distance considerations, even though other motivation to leave might be great. It could also be true that organizations would leave a desirable location (from a labor cost and raw materials perspective) to move to another location in which the marketing infrastructure to overseas markets is better. In short, a host of macrolevel constraints may complicate organizational decision making in this context.

The results regarding the subjective weights indicated that the weighted composite was significantly related ($r = .22$) to the actual location decision, but this correlation was relatively low and did not match that of the regression equation relating the eight climate dimensions to the relocation decision. There was some evidence that the two objective equations and the subject-

tively weighted equation yielded similar predicted values (matching indices were .46 and .44), but the rank orders of the weights presented in Table 5 indicated that the three sets of weights were fairly different. This result is, of course, consistent with much previous research on the coincidence of statistical and subjective weights (Schmitt & Levine, 1977).

In conclusion, there was no evidence that overall business climate perceptions were related to actual relocation decisions. There were modest relations between assessments of individual climate dimensions and overall perceptions of business climate, as well as the actual location decision. The absence of an attitude-behavior relation in this instance may be due to distance constraints on the businesses involved. The research suggests that business firms may have significant complaints about a particular environment that they cannot act on. However, given an appropriate opportunity, these firms, like individuals, may act in a manner consistent with their notions regarding business climate.

Future research on business climate would probably benefit from a consideration of differences among business types and from consideration of multiple dimensions of climate. The decision model used in this article should prove valuable, but additional attention should be given to the dependent variable. It may also be useful to examine the degree to which decision makers use nonlinear strategies to come to location decisions (Slovic, Fischhoff, & Lichtenstein, 1977; Tversky & Sattath, 1979) and to examine firms that are not constrained by distance considerations. These firms may be making decisions more consistent with their expressed attitudes.

As indicated earlier, the use of the lens model equations to study the aggregated judgments of a group of decision makers may have served to lower the various correlational indices. A similar combination of the judgments of poor, average, and good decision makers may serve to lower the validities of interview judgments. Research directed to a study of individual differences in the validity of interview judgments may be informative. The differences between decisions made by individuals, as opposed to groups of persons, should be investigated. For example, the business relocation decisions that were the subject of this article are likely made by a group of people after repeated deliberation, and the persons responding to our answer are more accurately described as reporters of these group discussions. We are not aware of studies of group decisions of this type within a policy-capturing framework. This type of study might serve as a bridge between macrolevel studies of organizational strategy and microstudies of individual decision making.

References

- Billings, R. S., & Marcus, S. A. (1983). Measures of compensatory and noncompensatory models of decision behavior: Process tracing versus policy capturing. *Organizational Behavior and Human Performance*, 31, 331-352.
- Brogden, H. E. (1946). On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Educational Psychology*, 37, 65-76.
- Cook, R. L., & Stewart, T. R. (1975). A comparison of seven methods for obtaining subjective descriptions of judgmental policy. *Organizational Behavior and Human Performance*, 13, 31-45.
- Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, 81, 95-106.
- Einhorn, H. J., Kleinmuntz, D. M., & Kleinmuntz, B. (1979). Linear regression and process tracing models of judgment. *Psychological Review*, 86, 465-485.
- Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.
- Hunt, T. L. (1985). *Michigan's business tax costs relative to the other Great Lakes states*. Kalamazoo, MI: W. E. Upjohn Institute for Employment Research.
- Mattila, J. M., & Thompson, W. R. (1983). *Study of studies re: Michigan business climate*. Report prepared for the Economic Alliance for Michigan. Detroit, MI: Wayne State University.
- Mobley, W. H., Horner, S. O., & Hollingsworth, A. T. (1978). An evaluation of precursors of hospital employee turnover. *Journal of Applied Psychology*, 63, 408-414.
- Nunnally, J. (1978). *Psychometric theory*. New York: McGraw-Hill.
- Schmitt, N. (1986). *Business relocation decisions in Michigan*. Report prepared for the Center for Redevelopment of Industrialized States. East Lansing: Michigan State University.
- Schmitt, N., & Levine, R. L. (1977). Statistical and subjective weights: Some problems and proposals. *Organizational Behavior and Human Performance*, 20, 15-30.
- Schmitt, N., Noe, R. A., & Gottschalk, R. (1986). Using the lens model to magnify raters' consistency, matching, and bias. *Academy of Management Journal*, 29, 130-138.
- Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Behavioral decision theory. *Annual Review of Psychology*, 28, 1-39.
- Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, 6, 649-744.
- Stumpf, S. A., & London, M. (1981). Capturing rater policies in evaluating candidates for promotion. *Academy of Management Journal*, 24, 752-766.
- Thompson, W. R., & Thompson, P. R. (1983). High-tech industries and high-tech places. *REI Review*, 1(1).
- Tversky, A., & Sattath, S. (1979). Preference trees. *Psychological Review*, 86, 542-573.
- U.S. Department of Commerce. (1984, June). *County of business patterns, 1982* (CBP-82-24). Washington, DC: Bureau of Census.
- Zedeck, S., & Kafry, D. (1977). Capturing rater policies for processing evaluation data. *Organizational Behavior and Human Performance*, 18, 269-294.

Received August 20, 1986

Revision received May 5, 1987

Accepted May 15, 1987 ■

Improving the Reliability of Eyewitness Identification: Putting Context Into Context

Brian L. Cutler and Steven D. Penrod
University of Wisconsin-Madison

Todd K. Martens
Harvard University

We examined the effects of context reinstatement procedures on eyewitness identification accuracy. Subjects were 290 undergraduates who viewed a videotaped reenactment of a liquor store robbery and, in a later session, attempted to identify the robber from a lineup parade. Two types of context reinstatement procedures were examined together with eight encoding, storage, and retrieval variables manipulated within the stimulus videotape and the lineup procedures. Disguise of the robber impaired identification accuracy ($p < .05$). There was a significant interaction between disguise and the context reinstatement interview ($p < .01$) such that the context reinstatement interview had a stronger impact on identification accuracy in the high-disguise condition. Lineup cues interacted with lineup composition ($p < .05$), retention interval ($p = .01$), and exposure to mug shots ($p = .05$; although in a manner contrary to our expectation). These interactions indicated that lineup context cues improved identification accuracy in the high-similarity, 2-week retention interval, and no mug-shots conditions.

The unreliability of eyewitness identification has been amply documented (Brigham, Maass, Snyder, & Spaulding, 1982; Clifford & Bull, 1978; Loftus, 1979; Penrod, Loftus, & Winkler, 1982; Yarmey, 1979). Experiments on eyewitness identification are typically designed to identify circumstances under which eyewitnesses are particularly fallible. Such research is valuable, as information gained from these experiments may enable police, prosecutors, defense attorneys, and juries to make more informed assessments of eyewitness identifications (Cutler, Penrod, & Martens, 1987; Cutler, Penrod, & Stuve, in press; Loftus, 1983; although see McCloskey & Egeth, 1983, for a dissenting view). It is perhaps equally important to identify procedures that might improve the reliability of eyewitness identifications. Despite the clear advantages to the criminal justice system of establishing such procedures, there has, until recently, been little research in this domain of eyewitness identification.

One promising approach to the task of improving the reliability of eyewitness identification and recall involves procedures designed to reinstate the context surrounding an event. According to the network theory of memory (Bower, 1981; Collins & Loftus, 1975), environmental, emotional, and other contextual and stimulus-relevant information are encoded into

memory, together with the to-be-remembered stimulus, as a set of nodes that are connected to the to-be-remembered stimulus through associative path links. Contextual cues to recognition prime alternative pathways to activate the node representing the to-be-remembered stimulus (henceforth referred to as the *stimulus node*). Sometimes an individual's information search fails to prime the necessary paths leading to the stimulus node (i.e., the individual fails to recall the necessary information). Under such circumstances, alternative pathways may be primed through the use of contextual cues, making it more probable that the stimulus node will be activated and the information in question recalled.

Given the theoretical rationale governing the effects of context cues, our research adopts a rather liberal definition of context. We define context as any information that is encoded together with the to-be-recognized stimulus. The information is presumed to be stored in the memory network and connected to the to-be-recognized stimulus through associative pathways (cf. Tulving & Thomson, 1973).

Although in theory, the network model should hold for both recall and recognition, the effects of context reinstatement procedures are noted for being stronger in tests of recall than in tests of recognition (Bower, 1981; Smith, in press). Consistent with this conclusion, experiments that have tested the effect of context reinstatement on identification accuracy have yielded mixed results. Some investigators (e.g., Krafka & Penrod, 1985; Malpass & Devine, 1981a) have shown positive effects for context reinstatement, whereas others (e.g., Cutler, Penrod, O'Rourke, & Martens, 1986; Davies & Milne, 1985; Sanders, 1984) have shown null or weak results for context reinstatement. It is difficult to precisely determine the locus of the differential effects for context reinstatement. Comparisons across experiments are made difficult due to differences in retention interval, type of incident, and context reinstatement procedure, all of which could conceivably mediate the influence

This research was funded by National Science Foundation Grant SES-8411721 and National Institute of Justice Grant 84-IJ-CX-0010 to the second author.

The authors wish to thank Carol Krafka and Peter Shapiro for their valuable assistance with the planning of this research, and Mark Bartells, Karla Bishell, James Coward, Michael Hart, Thomas O'Rourke, and Todd Ripple for their assistance with various phases of this research. We thank Roy Malpass and an anonymous reviewer for their comments concerning an earlier version of this article.

Correspondence concerning this article should be addressed to Brian L. Cutler, who is now at the Department of Psychology, Florida International University, North Miami Campus, North Miami, Florida 33181.

of context reinstatement procedures. Note that the mixed results for context reinstatement are unlikely to be attributable to Type I error, as a recent meta-analysis of the facial recognition literature (Shapiro & Penrod, 1986) reveals that, across a large number of studies, context cues improve identification accuracy.

Why might encoding, storage, and retrieval factors mediate the influence of context cues on eyewitness identification accuracy? One plausible answer is Smith's (in press) *outshining hypothesis*. The outshining hypothesis acknowledges that in a recognition test certain aspects of the stimulus itself serve as context cues to recognition. In Smith's terms, if information provided by the stimulus serves as a strong cue to recognition, then any additional incidentally encoded context cues are "outshined." In terms of the network theory, the cues provided by the stimulus itself are often sufficient to activate the necessary paths to the stimulus node, so additional attempts to activate alternative pathways to the stimulus node through the use of context cues yield few gains in recognition accuracy (see also Bower, 1981). If, on the other hand, the stimulus properties that ordinarily cue recognition are degraded or absent and, consequently, the appropriate pathway to the stimulus node is not activated, then additional context cues improve recognition accuracy by activating alternative pathways to the stimulus node. In the case of eyewitness identification, facial characteristics and hair qualities are known to be important cues. If these cues are optimally encoded, then other contextually reinstated cues may yield little or no improvement in identification accuracy. But if these important cues are degraded due to impoverished encoding (e.g., disguise, weapon focus), storage factors (e.g., exposure to mug shots, long retention interval), or retrieval conditions (e.g., variations in lineup procedures, suggestive instructions), then contextually reinstated cues may enhance the accuracy of identifications by facilitating the priming of alternative paths to the individual's memorial representation of the to-be-identified target.

In summary, the mixed results for context reinstatement procedures might be due to differences in the qualitative and quantitative nature of context reinstatement procedures used in previous research or to encoding, storage, and retrieval factors that for theoretical reasons might moderate the effectiveness of context reinstatement procedures. An experiment was carried out to examine the influence of two types of context reinstatement procedures and the interactions between these procedures and a variety of encoding, storage, and retrieval factors on eyewitness identification accuracy. Because our primary concern is with techniques that can be used for actual eyewitness situations, we limit our study of contextual cues to those that could conceivably be implemented within current police investigatory procedures.

A second important concern of the present research is the predictive validity of eyewitness confidence. It is often concluded that the correlation between confidence and identification accuracy is weak (Bothwell, Deffenbacher, & Brigham, 1986; Wells & Murray, 1984). Wells and Lindsay (1985) raised criticisms about the manner in which confidence has been assessed in the eyewitness studies; for example, they point out that confidence is generally measured with a single item and might therefore suffer from unreliability (although cf. Cutler, Penrod,

O'Rourke, & Marten, 1986; Cutler, Penrod, & Martens, 1987; Murray & Wells, 1982). We attempt to improve on earlier assessments by using multiple confidence indices. In addition, we assess both confidence in the ability to correctly identify the perpetrator (before subjects view the lineup), as well as confidence in the lineup judgment.

Overview of the Experiment

Subjects in the present experiment viewed a videotaped robbery of a liquor store, and later attempted to identify the robber from a lineup parade. In light of Smith's (in press) outshining hypothesis, we manipulated variables that have been shown to affect identification accuracy. Our general expectation was that context reinstatement procedures will be effective in improving identification accuracy in circumstances under which identification accuracy is less reliable due to factors such as disguises worn by the robber, mugshot searches, the presence of a weapon, and substantial retention interval. Two types of context reinstatement procedures were examined; one procedure involved prelineup interviews and the other involved exposing subject-witnesses to context cues embedded within the lineup itself.

Context Reinstatement Interview

The context reinstatement interview was modeled after the procedures used by Krafska and Penrod (1985) and by Cutler, Penrod, O'Rourke, & Martens (1986, Experiment 2). Like Krafska and Penrod, we attempted to reinstate context by using a guided interview consisting of mnemonic procedures ("mnemonic instructions") developed by Geiselman, Fisher, MacKinnon, and Holland (1985), and by exposing subjects to a series of snapshots depicting the victim of the robbery, as well as the environment in which the robbery occurred (snapshot review). The mnemonic instructions are comprised of the following procedures: (a) mental reinstatement of the context surrounding an incident, (b) report of all information recalled, (c) rehearsal of events in different orders, and (d) rehearsal of events from different perceptual perspectives. These mnemonic instructions alone have been shown to enhance the accuracy of eyewitness recall (Geiselman et al., 1985). The snapshot display consisted of photographs of the inside of the liquor store and of the clerk behind the counter. A third set of contextual cues was added to this procedure. Subjects were instructed to reread their written descriptions of the robbery and of the physical characteristics of the robber that they completed immediately after viewing the videotaped robbery (as in Cutler, Penrod, O'Rourke, & Martens, 1986, Experiment 2); this procedure is referred to as *original description review*. The mnemonic instructions, the snapshot review, and the original description review were combined into a single interview procedure and are henceforth referred to as the *context interview*.

Lineup Context Cues

The second type of context reinstatement procedures examined in the present experiment involved physical characteristic context cues tested by Cutler, Penrod, O'Rourke, & Martens

(1986, Experiment 1). In their experiment, subjects were systematically exposed to cues such as voice, gait, posture, skin color, and a three-fourths pose, and these cues were manipulated separately. In the present experiment, four of these cues—voice, gait, posture, and three-fourths pose—were combined into a single variable referred to as *lineup context cues* (all subjects were shown color lineups). Thus, one half of the subjects were shown a lineup consisting of snapshots and slides of a front and full profile of the head and shoulders only. The other half were shown the same lineup features, but were also shown the three-fourths pose and the full view of the suspects' bodies (posture cues) from the three poses. Subjects in this condition were also shown a videotaped segment of each lineup suspect walking in and out of the room in which the lineup was held (gait cues), and heard voice samples from each suspect, which consisted of a single spoken line (voice cues). Viewing time of each lineup suspect was, of course, held constant. Although these lineup features are somewhat different from traditional context reinstatement manipulations, they fit the broad definition of context cues specified earlier. Other researchers have also manipulated pose as a means to reinstating context (e.g., Thomson, Robertson, & Vogt, 1982).

In addition to the two types of context reinstatement manipulations, eight other encoding, storage, and retrieval variables were manipulated. These variables were chosen because they have been shown to affect identification accuracy in previous research. A description of each follows.

Disguise. In one half of the videotaped robberies, the robber wore a hat fully covering his hair, and in the other half, the robber wore no hat.

Weapon visibility. In one half of the videotaped robberies, the robber outwardly brandished his handgun during the entire robbery, whereas in the remaining versions, the robber's handgun remained hidden throughout the robbery.

Retention interval. Subjects attempted an identification after either 2 days or 2 weeks.

Exposure to mugshots. One half of the subjects in the present experiment searched a series of 41 mug-shot slides (which included neither the robber nor the other lineup members) for the robber during the encoding session of the experiment, and the other half viewed no mug shots.

Lineup instructions. Before viewing the lineup parade, one half of the subjects were given instructions that explicitly offered the option of rejecting the lineup. The remaining subjects were given instructions that failed to explicitly offer this option. Failing to offer the option of rejecting the lineup typically increases false identification (Cutler, Penrod, & Martens, 1987; Malpass & Devine, 1981b).

Lineup type. One half of the subjects attempted an identification from an offender-present lineup, whereas the remaining half attempted an identification from an offender-absent lineup.

Lineup size. One half of the subjects viewed 6-suspect lineups, and the other half viewed 12-suspect lineups.

Lineup composition. In the present experiment, lineups were created on the basis of similarity ratings obtained in pilot work; these lineups were intended to differ with respect to fairness. Subjects viewed either high-similarity lineups (lineups that contained several members who resembled the robber in appearance)

or low-similarity lineups (lineups that contained few members who resembled the robber in appearance).

Method

Subjects

Subjects ($N = 290$) were volunteers from the University of Wisconsin–Madison introductory psychology subject pool, who received extra credit points for their participation. Subjects were randomly assigned to conditions. There were from 2 to 4 subjects per cell. Each cell necessitated a separate encoding and retrieval session, and where possible, subjects within the same cell were run in groups.

Design

In all, 10 variables (two levels each) were manipulated within a $2^{(7+3)}$ fractional factorial design. A fractional factorial design (Kenny, 1985) is one in which some main effects are deliberately confounded with higher order interactions between other main effects. In the present experiment, 7 variables were fully crossed with one another in a 128-cell design, whereas another 3 variables were confounded with higher order four- and five-way interactions. The advantage to this design is that 10 main effects, 45 two-way interactions, and three-way interactions of interest can be meaningfully assessed within a $2^7 = 128$ -cell design. Testing 10 main effects and 45 two-way interactions in a full factorial design would necessitate $2^{10} = 1,024$ cells and many more subjects. The drawback to this fractional factorial design is that experimental efficiency is traded off against the fact that a few high-order interactions (four- and five-way) are confounded with main effects and there are many confounds among higher order interactions. Of course, these interactions tend to be uninterpretable in any event.

Materials

Stimulus videotapes. The plot of the vignette concerned a female clerk at a liquor store who, shortly after serving one customer, is confronted and robbed by a young man brandishing a handgun. The robber enters the store, demands the money from the register, and threatens to shoot the clerk. In the course of the interaction, the robber fires his weapon into the floor and roughs up the clerk. The entire videotape lasted approximately 100 s, and the robbery itself lasted approximately 75 s. Two variables, disguise and weapon visibility, were fully crossed within the videotapes. This was accomplished by a combination of repeated filming and editing of the robbery. Thus, in all, four videotapes were used. Stimulus videotapes were high quality, $\frac{3}{4}$ -in. videocassettes, and were shown on a large (64-in. diagonal) projector screen, using a Kloss Nova Beam, Model 2.

Instructions. All of the instructions were given in writing. Subjects in the mug-shot condition were instructed to study each of the mug shots and to search for the robber, while the experimenter read a number aloud for each mug shot. Immediately following the presentation of the mug shots, subjects were further instructed to indicate the number of the mug shot that they thought was the robber or indicate that the robber was not among the mug shots (unbiased instructions).

Subjects in the context interview condition first received written mnemonic instructions to think back through the event, from beginning to end, and then in different orders and from different perspectives. Subjects were also instructed to try to remember the emotions they felt during the robbery and recall everything they viewed. Subjects were given up to 5 min to reminisce.

Lineup materials. All of the subjects were given a set of color snapshots of each lineup suspect. Subjects studied a front view (face and shoulders) of each suspect on one side of the snapshot and a full profile

view on the reverse side of the snapshot. After subjects were given ample time (by their own acknowledgment) to examine the photo spread, subjects were shown a more detailed set of slides of each suspect. Only one suspect appeared on each slide. Subjects were allowed to study the photographs during the slide presentation in order to make comparisons among the suspects.

Subjects in the weak lineup context condition were shown slides of each suspect in front and full profile view, and slides consisted of views of the face and shoulders only. Subjects in the strong lineup context condition were shown slides of each suspect in front, three-fourth, and full profile views. In addition to views of the face and shoulders, full-body views of each suspect were also shown. Subjects in the strong lineup context condition also viewed a videotaped segment (¾ in. shown on a 25-in color monitor) of each lineup member walking in and out of the room in which the photographs and slides were taken, and heard a sample of each suspect’s voice. The voice sample consisted of a single line spoken about the weather, and each suspect spoke the same line. The presentation of the enhanced-lineup condition proceeded as follows: Subjects were shown a series of slides of a given suspect immediately followed by the videotaped segment. Voice samples were given during the videotaped segment. Presentation time of each suspect was held constant (35 s) across lineups.

Interrogation questionnaire. The description of the robber and the robbery, which all subjects completed but only one half of the subjects reread before seeing the lineup, was given in response to a series of 40 probes about the event. In addition, subjects completed a checklist containing nine categories of body build (e.g., slender, stocky), nine hair colors, nine hair styles (e.g., shoulder length, balding), six eye colors, seven facial hair categories, and nine overall descriptions (e.g., well kept, slouched, ugly). No restrictions were placed on the number of categories that subjects were allowed to check.

Confidence questionnaires. Confidence in lineup choice was assessed twice. Immediately after viewing the stimulus videotape during the encoding session, subjects completed a *prejudgment confidence* questionnaire that consisted of the following two questions: “If we showed you a lineup in which the robber was present, how confident are you that you could choose the right person?” and “If we showed you a lineup in which the robber was not present, how confident are you that you would not mistakenly choose somebody out of the lineup?” Responses to both inquiries were indicated on 9-point scales ranging from *not at all confident* (1) to *very confident* (9). Immediately after rendering a judgment on the lineup task, subjects completed a *postjudgment confidence* questionnaire that consisted of the following three questions: “How confident are you that your choice is correct?” “How willing are you to sign a sworn statement that your choice is correct?” and “What is the probability that your choice is correct?” Responses to the first two questions were given on 9-point scales ranging from *not at all confident* and *not at all willing* (1) to *very confident* and *very willing*, respectively (9). Responses to the third inquiry were open-ended and could theoretically range from 0 to 1.00.

Procedure

During the encoding session, subjects were first shown one of the four videotaped robberies. Subjects completed the prejudgment confidence questionnaire, and then completed a questionnaire in which they were asked to describe the robbery and the physical characteristics of the robber. Subjects who were not in the mug-shot condition were excused at this point. Subjects in the mug-shot condition were given the mug-shot instructions and were then shown the mug shots. After completing the mug-shot procedure, subjects were excused.

At the beginning of the retrieval session, subjects in the context reinstatement condition were administered the reinstatement procedures in the following order: mnemonic instructions, original description review,

and snapshot display. Subjects who were not in the context reinstatement condition were given an innocuous imagery assessment questionnaire. Subjects were then given lineup instructions (biased or neutral) and were handed the photographs of the lineup suspects. After subjects indicated that they had enough time to study the photographs, subjects were shown the lineup. At the completion of the lineup phase, subjects indicated their judgments privately and completed the postjudgment confidence questionnaire.

Results

Overall, 234 subjects (81%) made a positive identification, whereas 56 subjects (19%) rejected the lineup. Of the 140 subjects who were shown offender-present lineups, 90 (64%) correctly identified the robber, 37 (27%) mistakenly identified a foil, and 13 (9%) incorrectly rejected the lineup. Of the 150 subjects shown an offender-absent lineup, 43 (29%) correctly rejected the lineup, whereas 107 (71%) falsely identified a foil. Collapsed across lineup type, the overall correct performance rate (CP; proportion of hits + proportion of correct rejections) was .46.

Lineup Decisions

Malpass and Devine (1984) argued that when examining the effects of variables on identification accuracy, it is informative to first account for the variable’s effect on the lineup decision (i.e., the decision to identify a suspect or to reject the lineup), and through the consequence of that decision explain identification accuracy. This approach is especially appropriate if the effects of a variable on identification accuracy should theoretically be mediated through its effect on the lineup decision. Lineup instructions is the one factor in our experiment that meets this criterion. The remaining factors are hypothesized to affect identification accuracy by influencing sensitivity without necessarily affecting the lineup decision.

Thus, before analyzing the effects of lineup instructions on identification accuracy, we first analyzed its effects on the lineup decision. For the purpose of brevity, we limited our exploratory analyses of lineup decisions to main effects for each variable and the nine interactions between lineup instructions and each of the remaining factors. For this analysis, all positive identifications were scored 1 and lineup rejections were scored 0. A hierarchical regression analysis was then performed with each of the 10 factors entered on the first step and the 9 two-way interactions entered on the second step. The dependent variable

Table 1
Lineup Decisions: Predictors

Variable	M	d	t
Lineup type			
Offender absent	.72	.49	4.22*
Offender present	.90		
Lineup instructions			
Neutral	.69	.65	5.43*
Biased	.93		

Note. Means represent proportion of positive identifications.
* $p < .01$.

Table 2
Lineup Decisions: Summary Statistics

Step	R	Adjusted R ²	F total	df	MS _e	F change	df
1	.400	.130	5.33*	10, 279	.136		
2	.431	.129	3.25*	19, 270	.136	.95	9, 270

* $p < .01$.

was lineup decision (positive identification or lineup rejection). A summary of this regression analysis appears in Tables 1 and 2.

As expected, biased lineup instructions significantly increased the number of positive identifications. Of course, lineup type (offender present versus offender absent) also had a significant influence on the type of decision made. None of the other main effects or any of the two-way interactions were statistically significant.

Identification Performance

In order to examine the effects of the independent variables on identification performance, correct judgments were scored 1 and incorrect judgments were scored 0 (to form the CP score). With CP as the dependent variable, a hierarchical regression analysis was performed with the 10 predictor variables entered on the first step and the subsequent 45 two-way interactions entered on the second step. The 36 three-way interactions between lineup type (offender present vs. offender absent) and all other predictors were entered on the third step, to determine whether correct identifications and correct rejections were affected similarly by context and other variables. A summary of the regression results is displayed in Tables 3 and 4.

Significant main effects were found for lineup type and disguise of the robber. Fewer correct judgments were obtained in the offender-absent (as compared with offender-present) lineups and in the disguise (as compared with no-disguise) condition. Of the 45 two-way interactions, 6 were significant at $p < .05$, which is more than the number of significant interactions expected by chance alone.

Lineup context interacted significantly with lineup composition, such that context cues significantly improved performance if subjects were shown high-similarity lineups, but not if subjects were shown low-similarity lineups. Note that the difference in identification accuracy between the strong lineup cues–low similarity and strong lineup cues–high similarity cells (.54 vs. .44, respectively) is likely to be due to chance variation ($p > .05$).

Context cues in the lineup parade also interacted significantly with retention interval such that providing subjects with strong context cues in the lineup improved performance in the 2-week retention interval condition but had little effect on performance among subjects in the 2-day retrieval condition. Although it is apparent from the cell means that the strong cues in the lineup improved identification performance to a level above that obtained in the 2-day retention interval condition (.59 vs. .49), this difference is probably due to chance variation ($p > .05$).

The interaction between lineup context cues and exposure to mug shots was also significant. Contrary to our expectation, lineup context cues significantly improved identification accuracy among subjects who were not shown mug shots, but had little effect on identification performance among subjects who had been shown mug shots. The outshining hypothesis would lead us to expect the effects of context reinstatement to be stronger if subjects had been shown mug shots.

The context interview interacted significantly with disguise. It significantly improved identification performance if the robber had been disguised during the robbery, but had little effect on identification performance if the robber had not been disguised.

Disguise interacted significantly with lineup size such that disguise had a significantly larger effect on performance among subjects in the 12-suspect lineup condition than among subjects in the 6-suspect lineup condition.

Given that subjects in the biased lineup instruction condition have a strong tendency to make positive identifications (controlling for the presence of the offender in the lineup), biased lineup instructions should strongly reduce identification performance in the offender-absent condition, because any identification is an incorrect one. The effects of biased lineup instructions on identification performance in the offender-present condition should be less strong, because even though subjects are more inclined to make a positive identification, sometimes the identification is correct.

Biased instructions had a significantly larger impairment on identification performance in the offender-absent condition than in the offender-present conditions. This interaction might best be understood in terms of diagnosticity (Malpass & Devine, 1984; Wells & Lindsay, 1980). Diagnosticity refers to the ratio of correct identifications in offender-present lineups to false identifications in offender-absent lineups; therefore, higher ratios indicate better diagnosticity. Biased lineup instructions yielded a diagnosticity ratio of $.61/(1 - .13) = .70$, whereas neutral instructions yielded a diagnosticity ratio of $.67/(1 - .43) = 1.18$. Clearly, biased instructions strongly reduce the diagnosticity of identifications.

It is perplexing that the mug-shot manipulation had little effect on identification accuracy. One plausible hypothesis¹ for the lack of an effect is that the mug-shot search (in which the robber was *not* present) served to reinforce the subjects' beliefs that the robber might be absent from any additional set of mug shots or lineups that are to be searched. Such a hypothesis

¹ We are grateful to the anonymous reviewer for pointing out this plausible hypothesis.

Table 3
Identification Accuracy: Predictors

Variable	<i>M</i>	<i>d</i>	<i>t</i>
Step 1			
Lineup type			
Offender absent	.29	.77	6.52**
Offender present	.64		
Disguise			
Low	.51	-.24	-2.19*
High	.40		
Step 2			
Lineup Context × Lineup Composition			2.35*
Low Similarity			
Weak cues	.36	.40	3.06**
Strong cues	.54		
High Similarity			
Weak cues	.50	-.13	-0.99
Strong cues	.44		
Lineup Context × Retention Interval			2.91**
2-week retention interval			
Weak cues	.37	.49	3.75**
Strong cues	.59		
2-day retention interval			
Weak cues	.49	-.22	-1.68
Strong cues	.39		
Lineup Context × Exposure to Mug shots			-1.97*
No mug shots			
Weak cues	.35	.35	2.68**
Strong cues	.51		
Mug-shot search			
Weak cues	.51	-.09	-0.69
Strong cues	.47		
Context Interview × Disguise			2.87**
No disguise			
No interview	.57	-.22	-1.68
Interview	.47		
Disguise			
No interview	.29	.49	4.74**
Interview	.51		
Disguise × Lineup Size			-2.46**
12-suspect			
Low disguise	.51	-.31	-2.37**
High disguise	.37		
6-suspect			
Low disguise	.49	-.04	-0.31
High disguise	.47		
Lineup Instruction × Lineup Type			3.44**
Offender absent			
Neutral	.43	-.66	-5.04**
Biased	.13		
Offender present			
Neutral	.67	-.13	-0.99
Biased	.61		

Note. Means represent proportion correct.
* $p < .05$. ** $p < .01$.

would be supported by an interaction between lineup type (offender present vs. offender absent) and exposure to mug shots. The interaction, though, was nonsignificant.

The third step of the regression equation examined whether

correct identifications and correct rejections were affected similarly by context and other independent variables. This was accomplished by testing the three-way interactions between lineup (offender present vs. offender absent) and each of two-way interactions. The number of significant three-way interactions (3) did not exceed chance levels; these interactions are therefore not discussed. None of the significant two-way interactions were further qualified by three-way interactions with lineup type. Note, however, that 6 of 36 three-way interactions were not entered into the equation because of minimal tolerance—they were too highly correlated with other interaction terms. Given that the interactions with which these terms are confounded were nonsignificant, there is little threat to the interpretation of our results.

Confidence

Two subjects failed to complete the prejudgment confidence questionnaire, 4 subjects failed to complete two of the postjudgment confidence questions, and 14 subjects failed to complete one postjudgment confidence question; their data were therefore excluded from the following analyses. Mean prejudgment confidence regarding the ability to correctly identify the robber was 7.30 ($SD = 1.35$), and mean prejudgment confidence regarding the ability to correctly reject the offender-absent lineup was 6.02 ($SD = 1.73$). The average confidence rating was 5.76 ($SD = 2.00$) for postjudgment confidence in judgment on the lineup task, 4.06 ($SD = 2.41$) for willingness to sign a sworn statement, and .56 ($SD = .28$) for probability of a correct judgment. Correlations between confidence measures, identification performance, and choosing, are displayed in Table 5.

The two measures of prejudgment confidence correlated significantly, $r = .37$, $p < .01$, and the correlations between each of these measures and performance (CP) were of similar magnitude. Prejudgment confidence regarding the ability to correctly identify the suspect correlated .10 with performance, whereas prejudgment confidence regarding the ability to correctly reject the lineup correlated .07 with performance. The three postjudgment confidence ratings correlated highly with one another—the average correlation was .72, $p < .01$ —but minimally with prejudgment confidence measures—the average intercorrelation was .16, $p = .01$. Furthermore, the postjudgment confidence ratings correlated more highly with performance than did prejudgment confidence. For confidence in judgment, $r = .30$, $p < .01$, for willingness to sign a sworn statement, $r = .32$, $p < .01$, and for probability of correct judgment, $r = .27$, $p < .01$.

Clearly, pre- and postjudgment confidence are conceptually different measures; they demonstrate weak relations with one another, and they demonstrate different patterns of relations with identification performance. Given the internal consistency within these sets of measures, we decided to standardize and then aggregate prejudgment confidence measures to form a more reliable prejudgment confidence score and to standardize and then aggregate postjudgment confidence measures to form a more reliable postjudgment confidence score. The aggregate prejudgment confidence score (henceforth referred to as prejudgment confidence) correlated .10, $p = .05$, with performance, whereas the aggregate postjudgment confidence score

Table 4
Identification Accuracy: Summary Statistics

Step	R	Adjusted R ²	F total	df	MS _e	F change	df
1	.404	.134	5.47*	10, 279	.216		
2	.580	.180	2.15*	55, 234	.204	1.35	45, 234
3	.656	.193	1.81*	85, 204	.201	1.12	30, 204

* $p < .01$.

(henceforth referred to as postjudgment confidence) correlated .33, $p < .01$, with performance. The correlation between pre- and postjudgment confidence was .22, $p < .01$.

The estimator and system variables were also tested for their effects on confidence and as moderators of the confidence-accuracy correlation. The first analysis examined the aggregate pre-judgment confidence as the dependent measure. Identification accuracy was entered on the first step, disguise and weapon visibility on the second, and the two interaction terms (Disguise \times Identification Accuracy and Weapon Visibility \times Identification Accuracy) were entered on the third step. The interaction terms test heterogeneity of variance, or the moderator effects. Disguise was a significant predictor of prejudgment confidence (semipartial $r = -.17$, $p < .01$). No other predictors were significant.

A similar analysis was then carried out with the aggregate postjudgment confidence as the dependent variable. Identification accuracy was entered on the first step, the 10 predictors on the second, and the 10 interactions between identification accuracy and each predictor were entered on the third step. Exposure to mug shots significantly reduced postjudgment confidence (semipartial $r = -.17$, $p < .01$). Subjects shown an offender-present lineup were significantly more confident than were subjects shown an offender-absent lineup (semipartial $r = .14$, $p < .05$). None of the moderator effects were statistically significant.

Discussion

Results of this experiment indicate that both context reinstatement interviews and lineup context cues affect identifica-

tion performance by interacting with other variables that affect identification performance. The procedures differ, however, in the variables with which they interact. The effectiveness of the context interview was mediated by the disguise of the robber. This disguise, which consists of a hat covering most of the robber's hair, reliably reduces identification accuracy (Cutler, Penrod, O'Rourke, & Martens, 1986; Cutler, Penrod, & Martens, 1987; Shapiro & Penrod, 1986). The context reinstatement interview was more effective in the disguise condition (poor encoding conditions) than in the no-disguise condition.

The effects of lineup contextual cues were moderated by the similarity of the lineup members to the target, retention interval, and exposure to mug shots. Lineup context cues improved identification performance if subjects were presented with high-similarity lineups; that is, if lineup suspects resembled the offender in physical appearance. It is generally agreed that lineups should be fair tests (Doob & Kirshenbaum, 1973; Malpass & Devine, 1983; Wells, Leippe, & Ostrom, 1979) and should therefore contain foils who look like the suspect. Results of this experiment show that when lineups do contain such foils, physical characteristic context cues contribute to improved identification performance.

Retention interval is a strong predictor of recognition accuracy in facial recognition and eyewitness identification experiments (Shapiro & Penrod, 1986). As shown in the present experiment, the effectiveness of context reinstatement varied as a function of retention interval, thus complicating comparisons of context effects across experiments. This finding is especially noteworthy given that in the eyewitness identification literature, retention intervals vary from less than 1 hour (e.g., Sanders, 1984) to 5 months (Malpass & Devine, 1981a).

Table 5
Correlations Between Confidence and Identification Accuracy

Measure	1	2	3	4	5	6	7	8	9
Prejudgment confidence									
1. Correct identification	—	.37	.83	.24	.23	.16	.23	.10	-.01
2. Correct rejections		—	.82	.14	.11	.11	.13	.07	.03
3. Aggregate			—	.23	.21	.16	.22	.10	.01
Postjudgment confidence									
4. Decision specific				—	.80	.70	.92	.30	-.04
5. Willingness to sign a sworn statement					—	.66	.91	.27	-.07
6. Probability						—	.87	.27	.01
7. Aggregate							—	.33	-.04
Identification performance									
8. CP								—	-.30
9. Choosing									—

Note. $N = 274$. Correlations above .10 are significant at $p < .05$; correlations above .14 are significant at $p < .01$. CP = identification accuracy.

Lineup context cues improved identification performance among subjects who were not shown mug shots, but had less of an effect among subjects who were shown mug shots. This finding is perplexing, as it is opposite to the hypothesis derived from the outshining hypothesis.

With respect to eyewitness confidence, our results are consistent with those obtained in previous experiments (Cutler, Penrod, O'Rourke, & Martens, 1986; Cutler, Penrod, & Martens, 1987). In general, the confidence-performance relation was found to be significant, but moderate in size, and confidence in lineup decision was a much better predictor of identification performance than was confidence in ability to identify the offender. The finding that prejudgment confidence correlated so weakly with identification accuracy suggests that a witness's initial confidence in his or her ability to identify a perpetrator should not be used to support the validity of an identification. In addition, the finding implies that witnesses whose confidence in their ability to correctly identify a perpetrator wanes, might nevertheless be encouraged to attempt an identification.

Although the amount of variance in performance accounted for by confidence is not great, it is important to note that (a) as illustrated elsewhere (Cutler & Penrod, 1986; Nunnally, 1978), the point-biserial correlation is highly susceptible to attenuation due to differentially skewed distributions among the dichotomous variable (identification accuracy) and the continuous variable (confidence), and (b) the absolute value of a correlation coefficient may be a better estimate of strength of association than is the squared correlation coefficient (Ozer, 1985; see also Wells & Lindsay, 1985).

Together with the results of the Krafska and Penrod (1985) and Malpass and Devine (1981a) studies, the results of the present experiment should help to dispel some pessimistic notions regarding the effects of contextual cues on identification accuracy. We have found that physical characteristic cues in the lineup and reinstatement of environmental and emotional state context are effective in enhancing identification performance. However, our results, together with the other experiments reviewed, suggest that there must be an appreciable retention interval or impairment of memory due to other encoding, storage, or retrieval factors for context reinstatement to be effective. Physical characteristic context cues appear to be especially sensitive to retention interval. From a theoretical perspective, it would be helpful to know why some context cues are more effective than other context cues and why various context cues are differentially mediated by encoding, storage, and retrieval factors. From a forensic perspective, more research is needed to identify procedures that increase the reliability of eyewitness memory—in conjunction, naturally, with research that identifies conditions under which eyewitness memory is likely to be fallible. Reinstatement of context is a promising approach to the enhancement of eyewitness performance—one that surely deserves further attention in both applied and theoretical realms.

References

- Bothwell, R. K., Deffenbacher, K. A., & Brigham, J. C. (1986). *Correlation of eyewitness accuracy and confidence: The optimality hypothesis revisited*. Manuscript submitted for publication.
- Bower, G. H. (1981). Mood and memory. *American Psychologist*, 36, 129-148.
- Brigham, J. C., Maass, A., Snyder, L. D., & Spaulding, K. (1982). Accuracy of eyewitness identifications in a field setting. *Journal of Personality and Social Psychology*, 42, 673-680.
- Clifford, B. R., & Bull, R. (1978). *The psychology of person identification*. London: Routledge & Kegan Paul.
- Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 5, 85-88.
- Cutler, B. L., & Penrod, S. D. (1986, May). *Eyewitness calibration: A within-subject perspective on the eyewitness confidence-accuracy relation*. Paper presented at the meeting of the Midwestern Psychological Association, Chicago.
- Cutler, B. L., Penrod, S. D., & Martens, T. K. (1987). The reliability of eyewitness identifications: The role of system and estimator variables. *Law and Human Behavior*, 11, 233-258.
- Cutler, B. L., Penrod, S., O'Rourke, T. E., & Martens, T. K. (1986). Unconfounding the effects of contextual cues on eyewitness identification accuracy. *Social Behavior*, 1, 113-134.
- Cutler, B. L., Penrod, S. D., & Stuve, T. E. (in press). Juror decisionmaking in eyewitness identification cases. *Law and Human Behavior*.
- Davies, G. M., & Milne, A. (1985). Eyewitness composite production as a function of mental or physical reinstatement of context. *Criminal Justice and Behavior*, 12, 209-220.
- Doob, A. N., & Kirshenbaum, H. M. (1973). Bias in police lineups: Partial remembering. *Journal of Police Science and Administration*, 1, 287-293.
- Geiselman, R. E., Fisher, R. P., MacKinnon, D. P., & Holland, H. L. (1985). Eyewitness memory enhancement in the police interview: Cognitive retrieval mnemonics versus hypnosis. *Journal of Applied Psychology*, 70, 401-412.
- Kenny, D. (1985). Quantitative methods for social psychology. In G. Lindzey & E. Aronson (Eds.), *The handbook for social psychology* (Vol. 1, pp. 487-508). Hillsdale, NJ: Erlbaum.
- Krafska, C., & Penrod, S. (1985). Reinstatement of context in a field experiment on eyewitness identification. *Journal of Personality and Social Psychology*, 49, 58-69.
- Loftus, E. F. (1979). *Eyewitness testimony*. Cambridge, MA: Harvard University Press.
- Loftus, E. F. (1983). Silence is not golden. *American Psychologist*, 38, 564-572.
- Malpass, R. S., & Devine, P. G. (1981a). Guided memory in eyewitness identification. *Journal of Applied Psychology*, 66, 343-350.
- Malpass, R. S., & Devine, P. G. (1981b). Eyewitness identification: Lineup instructions and the absence of the offender. *Journal of Applied Psychology*, 66, 482-489.
- Malpass, R. S., & Devine, P. G. (1983). Measuring the fairness of eyewitness identification lineups. In S. M. A. Lloyd-Bostock & B. R. Clifford (Eds.), *Evaluating witness evidence* (pp. 81-102). New York: Wiley.
- Malpass, R. S., & Devine, P. G. (1984). Research on suggestion in lineups and photospreads. In G. L. Wells & E. F. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives* (pp. 64-91). New York: Cambridge University Press.
- McCloskey, M., & Egeth, H. (1983). Eyewitness identification: What can a psychologist tell a jury? *American Psychologist*, 38, 550-563.
- Murray, D. M., & Wells, G. L. (1982). Does knowledge that a crime was staged affect eyewitness accuracy? *Journal of Applied Social Psychology*, 12, 42-53.
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
- Ozer, D. J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin*, 97, 307-315.
- Penrod, S., Loftus, E. F., & Winkler, J. D. (1982). Eyewitness reliability.

- In R. Bray & N. Kerr (Eds.), *The psychology of the courtroom* (pp. 119–168). New York: Academic Press.
- Sanders, G. L. (1984). The effects of context cues on eyewitness identification responses. *Journal of Applied Social Psychology*, 14, 386–397.
- Shapiro, P., & Penrod, S. D. (1986). A meta-analysis of the facial recognition studies. *Psychological Bulletin*, 100, 139–156.
- Smith. (in press). Environment context-dependent memory. In G. Davies & D. Thomson (Eds.), *Memory in context: Context in memory*. Chichester, England: Wiley.
- Thomson, D. M., Robertson, S. L., & Vogt, R. (1982). Person recognition: The effect of context. *Human Learning*, 1, 137–154.
- Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80, 352–373.
- Wells, G. L., Leippe, M. R., & Ostrom, T. M. (1979). Guidelines for empirically assessing the fairness of a lineup. *Law and Human Behavior*, 3, 285–293.
- Wells, G. L., & Lindsay, R. C. L. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin*, 88, 776–786.
- Wells, G. L., & Lindsay, R. C. L. (1985). Methodological notes on the confidence–accuracy relationship in eyewitness identification. *Journal of Applied Psychology*, 70, 413–419.
- Wells, G. L., & Murray, D. M. (1984). Eyewitness confidence. In G. L. Wells & E. F. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives* (pp. 155–170). New York: Cambridge University Press.
- Yarmey, A. D. (1979). *The psychology of eyewitness testimony*. New York: Free Press.

Received June 25, 1986

Revision received April 15, 1987

Accepted April 16, 1987 ■

Two (or More?) Dimensions of Organizational Commitment: Reexamination of the Affective and Continuance Commitment Scales

Gail W. McGee and Robert C. Ford

Department of Management, University of Alabama at Birmingham

Two distinct views of organizational commitment have developed, one that regards it as attitudinal and the other as behavioral. Meyer and Allen (1984) acknowledged the importance of both approaches (labeling them affective and continuance commitment) and developed scales for measuring each. The present study reexamined some psychometric properties of these scales. The affective commitment scale appeared to be unidimensional and had good internal consistency reliability. For the continuance commitment scale, however, two distinct dimensions were identified. The first reflected commitment based on few existing employment alternatives, and the second reflected commitment based on personal sacrifice associated with leaving the organization. Affective commitment was correlated significantly and negatively with the first dimension (low alternatives) and significantly and positively with the second dimension (high personal sacrifice). Recommendations for future use of these scales are discussed.

In recent years many studies have focused on the concept of organizational commitment. One literature review (Reichers, 1985) identified 11 studies treating commitment as an independent variable and more than 20 viewing it as a dependent variable. Despite so much attention, the concept itself (and, therefore, its measurement) is not thoroughly understood.

Two views of commitment have dominated the literature. The first sees commitment as affective or attitudinal. The individual identifies with the organization and, therefore, is committed to maintain membership in order to pursue its goals. This approach typically has been operationalized with a scale developed by Porter, Steers, Mowday, and Boulian (1974).

A quite different view of organizational commitment known as the “side-bet theory” evolved from the work of Becker (1960), who regarded commitment as behavioral rather than attitudinal. According to this view, the individual is bound to the organization through extraneous interests (e.g., pensions, seniority) rather than favorable affect toward the organization. Behavioral commitment usually has been operationalized with scales developed by Ritzer and Trice (1969) and modified by Hrebiniak and Alutto (1972).

Meyer and Allen (1984), who labeled these two views as affective and continuance commitment, suggested that they have been confounded in previous studies. Meyer and Allen stated that “the measure used to test Becker’s side-bet theory of commitment is saturated with affective commitment and, as such, does not allow the theory to be tested appropriately” (p.

378). They developed and tested alternative scales to measure both affective and continuance commitment. The purpose of this study was to reexamine some psychometric properties of the Meyer and Allen scales.

Method

Subjects and Procedure

Questionnaires were mailed to 997 faculty from 4-year colleges and universities in the United States and Canada whose names were selected randomly from the *National Faculty Directory* (1984). Twenty-six subjects could not be contacted because of retirement, relocation, or death. Of the remaining 971 subjects, 350 responded for a return rate of 36%. According to Locke, Fitzpatrick, and White (1983), this percentage is within the normal range for faculty surveys. Comparing respondents with nonrespondents on available demographic characteristics indicated no major differences.

Measures

Affective commitment was measured using Meyer and Allen’s (1984) eight-item Affective Commitment Scale (ACS). Continuance commitment also was measured using an eight-item Continuance Commitment Scale (CCS) constructed by Meyer and Allen. In two samples, Meyer and Allen reported internal consistency reliability estimates (Cronbach’s alpha) of .88 and .84 for the ACS and .73 and .74 for the CCS.

Results

Demographic Data

Subjects represented a broad range of ages, disciplinary areas, geographic regions, institutional sizes, and organizational tenure. Ages ranged from 27 years to 71 years (M age = 48). Tenure ranged from 1 to 40 years; average tenure was 14.5 years. Demographic characteristics of our subjects were compared

We wish to thank the editor for his helpful suggestions on an earlier draft of this article. We also gratefully acknowledge the comments of two anonymous reviewers.

Correspondence concerning this article should be addressed to Gail W. McGee, Department of Management, University of Alabama at Birmingham, University Station, Birmingham, Alabama 35294.

Table 1
Factor Analysis of 16 Organizational Commitment Items: Two-Factor Solution

Item		Factor loadings	
		Factor 1	Factor 2
ACS1:	I do not feel a strong sense of belonging to my organization.	.899	-.221
ACS2:	I do not feel "emotionally attached" to this organization.	.832	-.139
ACS3:	This organization has a great deal of personal meaning for me.	.818	-.059
ACS4:	I do not feel like "part of the family" at this organization.	.707	-.162
ACS5:	I would be very happy to spend the rest of my career with this organization.	.662	.146
ACS6:	I enjoy discussing my organization with people outside it.	.618	-.045
ACS7:	I really feel as if this organization's problems are my own.	.557	.040
ACS8:	I think I could easily become as attached to another organization as I am to this one.	.419	.133
CCS1:	Right now, staying with my organization is a matter of necessity as much as desire.	-.139	.624
CCS2:	One of the major reasons I continue to work for this organization is that leaving would require considerable personal sacrifice—another organization may not match the overall benefits I have.	.247	.618
CCS3:	I feel I have too few options to consider leaving this organization.	-.110	.586
CCS4:	One of the few negative consequences of leaving this organization would be the scarcity of available alternatives.	-.196	.548
CCS5:	It would be very hard for me to leave my organization right now, even if I wanted to.	.385	.534
CCS6:	Too much in my life would be disrupted if I decided I wanted to leave my organization now.	.289	.518
CCS7:	It wouldn't be too costly for me to leave my organization in the near future.	.167	.198
CCS8:	I am not afraid of what might happen if I quit my job without having another one lined up.	-.071	.164

Note. ACS = Affective Commitment Scale. CCS = Continuance Commitment Scale. Items were presented in random order in the questionnaire; reverse-worded items were reverse scored prior to data analyses.

with corresponding national data from the *Fact Book for Academic Administrators* (Anderson, 1981). Seventy-eight percent of our subjects were men (73% nationally), and 75% were associate or full professors (50% nationally). Twenty percent of our subjects were in the social sciences (20% nationally), 13% in the physical sciences (13% nationally), 6% in engineering (8% nationally), 22% in the humanities (25% nationally), 10% in the biological sciences (7% nationally), 8% in business (5% nationally), and 11% in education (13% nationally). Eleven percent of our subjects were from other areas (e.g., social work, home economics). Our subjects, then, appeared to adequately represent the U.S. population of college and university faculty.

Psychometric Properties of the Commitment Scales

Dimensionality of the scales. The 16 items comprising the two scales were factor analyzed using maximum likelihood estimation followed by varimax rotation. Two factor analyses were performed: the first specifying two factors (as suggested by Meyer & Allen, 1984) and the second forcing no specific number of factors. The results of these analyses are shown in Tables 1 and 2.

In the two-factor solution the eight ACS items loaded on the first factor. Six of the eight continuance-commitment items loaded strongly on the second factor. These results generally supported the scale dimensionality suggested by Meyer and Allen (1984), although two CCS items did not have acceptable loadings.

The second analysis yielded four factors, three of which were interpretable. Seven of the eight affective-commitment items clearly loaded on the first factor. Examination of the item word-

ing for the second and third factors revealed that the three items that constituted the second factor reflected the role of available alternatives in the decision to remain on one's job; the items loading on the third factor reflected personal sacrifice that would result from leaving the organization. Meyer and Allen

Table 2
Factor Analysis of 16 Organizational Commitment Items: Four-Factor Solution

Item	Factor loadings			
	Factor 1	Factor 2	Factor 3	Factor 4
ACS1	.894	-.224	.085	-.004
ACS3	.835	-.019	.082	.007
ACS2	.832	-.130	.096	-.005
ACS4	.706	-.154	.070	-.054
ACS6	.572	-.058	.235	.051
ACS5	.572	-.052	.409	-.004
ACS7	.537	.008	.164	-.053
ACS8	.333	-.051	.333	.081
CCS4	-.159	.689	.058	.002
CCS3	-.084	.686	.099	.130
CCS1	-.166	.580	.241	.089
CCS6	.129	.177	.670	.018
CCS5	.241	.221	.631	.060
CCS2	.152	.408	.475	.086
CCS7	.117	-.015	.157	.980
CCS8	-.069	.146	-.004	.293

Note. ACS = Affective Commitment Scale. CCS = Continuance Commitment Scale. Item numbers correspond to numbers shown in Table 1. Highest loadings are underscored.

Table 3
Interscale Correlations For Affective and Continuanace Commitment

Scale	1	2	3	4
1. Affective Commitment (8 items)	—			
2. Continuanace Commitment (8 items)	.08	—		
3. Continuanace Commitment: Low alternatives (3 items)	-.21*	.76*	—	
4. Continuanace Commitment: High sacrifice (3 items)	.34*	.78*	.37*	—

* *p* < .001.

(1984) stated that they experimentally manipulated subjects' perceptions of continuance commitment "by providing information about the investments (side bets) the individual had accumulated and the extent to which he perceived alternatives to his present job" (p. 374). Meyer and Allen actually described the two different aspects or dimensions of commitment reflected by the second and third factors in the present study.

Reliabilities of the scales. The factor analyses suggested deleting two CCS items and recomputing the remaining six items as two subscales. Internal consistency reliability estimates (Cronbach's alpha) were calculated for the following four scales: (a) the original ACS, (b) the original CCS, (c) the new subscale CCS (Low Perceived Alternatives [CC:LowAlt]), and (d) the new subscale CCS (High Personal Sacrifice [CC:HiSac]). Reliability coefficients for these scales were .88, .70, .72, and .71, respectively. Because the two new CC subscales contained only three items each, the internal consistency reliability estimates of .72 and .71 were quite acceptable. In fact, they were slightly higher than the internal consistency estimate for the original CCS, despite its greater number of items. The development of additional items for each subscale might improve their reliabilities and bring them into line with the ACS.

Correlations among commitment scales. Correlations among the ACS, CCS, CC:HiSac, and CC:LoAlt subscales are shown in Table 3. In one sample, Meyer and Allen (1984) reported a correlation between the ACS and CCS of $-.01$; in a second sample, the correlation was .25, which was reported to be insignificant. The correlation of .08 between the ACS and CCS in our study supports their finding. However, examination of the subscale correlations suggests a different conclusion. Affective commitment was significantly and negatively related to CC:LowAlt, but significantly and positively related to CC:HiSac. Employees who were more emotionally committed were significantly less likely to remain because of a perceived lack of alternatives, but significantly more likely to perceive great personal sacrifice related to leaving. The apparent lack of association between the ACS and CCS occurred because the correlations between the ACS and the two subscales were approximately equal in magnitude but opposite in sign.

Discussion

This study lends substantial support to the conceptual distinction between affective and continuance commitment (Meyer & Allen, 1984; Reichers, 1985; Steers & Porter, 1983). Furthermore, the study provides additional evidence for the usefulness of the commitment scales developed by Meyer and Allen. However, our results suggest some caution when using the scales.

The ACS, as developed by Meyer and Allen (1984), had good internal consistency reliability and was unidimensional. The CCS, however, was not unitary. It consisted of two unique components: the first based on perceptions that few employment alternatives exist and the second on high personal sacrifice associated with leaving the organization. The second subscale, CC:HiSac, appears to more closely parallel the side-bet view of commitment, as described originally by Becker (1960). Some people may indeed remain with an organization simply because they perceive no alternatives, but we believe that this reflects a quite different form of "commitment" than either the affective or side-bet view. Thus, it might be useful in the future to develop further this two-pronged approach to behavioral commitment. The development of additional items, similar to those that constitute CC:HiSac, could strengthen and refine the scale, making it more useful for testing the side-bet theory of commitment. Similarly, a greater understanding of organizational continuity based on a lack of alternatives could be gained by developing additional items for the CC:LoAlt scale.

Meyer and Allen (1984) reported that continuance commitment and affective commitment were unrelated. In contrast, we found that the two CC subscales were significantly, though differentially, related to affective commitment. Steers and Porter (1983) stated that the "commitment process may be viewed as a self-reinforcing cycle in which attitudes and behaviors are reciprocally related" (p. 449). Additionally, they suggested that individuals who feel bound to an organization (through side bets or sunk costs) "... typically engage in some form of psychological bolstering in which they attempt to rationalize or self-justify their situation. . . ." (p. 428). Thus, a high degree of behavioral commitment could produce affective commitment through a process of dissonance reduction. Our results do not explain the nature of the association, but they do suggest that affective and continuance commitment may not operate totally independently of one another.

If future studies offer additional confirmation of the multidimensional nature of organizational commitment, the ability to clarify the contradictory and inconclusive relations between commitment and its antecedents and consequents will be considerably improved. The additional insight gained through this study into the dimensionality of the construct hopefully will be a useful step toward this clarification.

References

Anderson, C. J. (1981). *Fact book for academic administrators: 1981-82*. Washington, DC: American Council on Education.
Becker, H. S. (1960). Notes on the concept of commitment. *American Journal of Sociology*, 66, 32-42.

- Hrebiniak, L. G., & Alutto, J. A. (1972). Personal and role-related factors in the development of organizational commitment. *Administrative Science Quarterly*, 17, 555-572.
- Locke, E. A., Fitzpatrick, W., & White, F. M. (1983). Job satisfaction and role clarity among university and college faculty. *The Review of Higher Education*, 6, 343-365.
- Meyer, J. P., & Allen, N. J. (1984). Testing the "side-bet theory" of organizational commitment: Some methodological considerations. *Journal of Applied Psychology*, 69, 372-378.
- National Faculty Directory: 1985 (1984). (15th ed.; Vols. 1-3). Detroit, MI: Gale.
- Porter, L. W., Steers, R. M., Mowday, R. T., & Boulian, P. V. (1974). Organizational commitment, job satisfaction, and turnover among psychiatric technicians. *Journal of Applied Psychology*, 59, 603-609.
- Reichers, A. E. (1985). A review and reconceptualization of organizational commitment. *Academy of Management Review*, 10, 465-476.
- Ritzer, G., & Trice, H. M. (1969). An empirical study of Howard Becker's side-bet theory. *Social Forces*, 47, 475-479.
- Steers, R. M., & Porter, L. W. (1983). Employee commitment to organizations. In R. M. Steers & L. W. Porter, *Motivation and work behavior* (pp. 441-451). New York: McGraw-Hill.

Received December 16, 1985

Revision received February 16, 1987

Accepted January 26, 1987 ■

Instructions to Authors

Articles submitted for publication in the *Journal of Applied Psychology* are evaluated according to the following criteria: (a) significance of contribution, (b) technical adequacy, (c) appropriateness for the journal, and (d) clarity of presentation. In addition, articles must be clearly written in concise and unambiguous language. They must be logically organized, progressing from a statement of problem or purpose, through analysis of evidence, to conclusions and implications.

Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (3rd ed.). Articles not prepared according to the guidelines of the *Manual* will not be reviewed. All manuscripts must include an abstract of 100-150 words typed on a separate sheet of paper. Typing instructions (all copy must be double-spaced) and instructions on preparing tables, figures, references, metrics, and abstracts appear in the *Manual*. Also, all manuscripts are subject to editing for sexist language.

Authors can refer to recent issues of the journal for approximate length of regular articles. (Four double-spaced manuscript pages equal one printed page.) A few longer articles of special significance are occasionally published as monographs. Short Notes feature brief reports on studies such as those involving some methodological contribution or important replication.

APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more journals. APA policy also prohibits duplicate publication, that is, publication of a manuscript that has already been published in whole or in substantial part in another journal. Also, authors of manuscripts submitted to APA journals are expected to have available their raw data throughout the editorial review process and for at least 5 years after the date of publication.

Authors will be required to state in writing that they have complied with APA ethical standards in the treatment of their sample, human or animal, or to describe the details of treatment. (A copy of the APA Ethical Principles may be obtained from the APA Ethics Office, 1200 17th Street, N.W., Washington, DC 20036.)

Anonymous reviews are optional, and authors who wish anonymous reviews must specifically request them when submitting their manuscripts. Each copy of a manuscript to be anonymously reviewed should include a separate title page with authors' names and affiliations, and these should not appear anywhere else on the manuscript. Footnotes that identify the authors should be typed on a separate page. Authors should make every effort to see that the manuscript itself contains no clues to their identities.

Manuscripts should be submitted in quadruplicate and all the copies should be clear, readable, and on paper of good quality. A dot matrix or unusual typeface is acceptable only if it is clear and legible. Authors should keep a copy of the manuscript to guard against loss. Mail manuscripts to the Editor, Robert Guion, Department of Psychology, Bowling Green State University, Bowling Green, Ohio 43403.

Some Time Dimensions of Work: Measurement of an Underlying Aspect of Organization Culture

Jacquelyn B. Schriber

California School of Professional Psychology—Los Angeles

Barbara A. Gutek

The Claremont Graduate School

We examined the existence of temporal dimensions of organization culture (e.g., norms of time in organizations) and developed an instrument to measure those dimensions to facilitate cross-organizational and intraorganizational comparisons. A questionnaire designed to measure 15 hypothesized temporal dimensions was completed by 529 respondents from 51 work groups in 23 organizations. The sample was selected to meet certain criteria concerning organization type, organization size, and work group type. A principal components analysis extracted 13 usable scales: Time Boundaries Between Work and Nonwork, Sequencing of Tasks, Punctuality, Allocation, Awareness, Synchronization and Coordination, Variety Versus Routine, Intraorganizational Time Boundaries, Future Orientation, Schedules and Deadlines, Work Pace, Autonomy of Time Use, and Quality Versus Speed. Two hypothesized scales (Buffer in Planning and Buffer in Workday) did not emerge. Applications of the Time Dimensions Scales are discussed.

Time is a basic dimension of organizations. How time is partitioned, scheduled, and used has both dramatic and subtle influences on organizations and the people in them. For organizations, the effective scheduling, coordination, and synchronization of people and tasks through time is a key to survival, growth, and profitability. For employees, the effective scheduling and use of time across tasks both at work and outside work can affect their performance and satisfaction, on the job and off. In addition, understanding the norms about time at work (e.g., conforming to schedules, deadlines, work pace) can spell the difference between an employee's success or failure within a work organization.

The present exploratory study was designed to develop measures of temporal dimensions of work that can be used in research on time in organizations. As background, we discuss the relation of time to organization culture and define several aspects of time applicable to organizational behavior.

Relation of Time to Organization Culture

Norms about time can be viewed as characteristics of culture (Schriber, 1985). The cultures of western, developed countries generally hold views of time that are substantially different from those associated with cultures in less developed countries. Similarly, it has been noted that organizations have different cultures (e.g., Gregory, 1983; Jelinek, Smircich, & Hirsch, 1983; Jones, 1983; Martin & Siehl, 1983; Smircich, 1983; Wilkins & Ouchi, 1983). Differences among organization cultures should lead to differences among views and assumptions about time

among organizations (Schein, 1983). In addition, views or norms about time guide behavior (Doob, 1971).

Norms have temporal components that help integrate complex work processes and thereby facilitate the flow of work. For example, Zerubavel (1981, p. 3) contended that norms prescribe "sequential rigidity" and Doob (1971) viewed norms as facilitating conformity. Rigidity and conformity to the temporal aspects of norms provide necessary behavior control when complex, multiple goals can only be attained through coordinated activity. McGrath and Rotchford (1983) further noted that not all actions can be tightly specified. Norms and their temporal aspects guide behavior to increase individual rewards and decrease costs associated with interactions.

Although the temporal and content aspects of norms work to homogenize individuals' behaviors within groups, different groups have different norms—many of which may differ in their temporal aspects. Doob (1971) suggested that persons regulate their behavior to conform to groups to which they belong or would like to belong, or to those which have some power over them. Customs and norms concerning the uses and meanings of time (acquired through socialization) help indicate group membership, identify members' values, and explain patterns of behavior (e.g., whether people are precisely on time for meetings or whether meetings habitually start later than scheduled; Doob, 1971). People may alter their uses of time, as well as their notions about appropriate time norms and values, when they enter a particular organization (e.g., work organization, social club). Organizations and their subunits, each with different throughput and process priorities, may have different norms about time.

This article is based on a doctoral dissertation completed at The Claremont Graduate School by the first author.

We would like to thank Dale Berger for his statistical advice and Stuart Oskamp for his helpful suggestions.

Correspondence concerning this article should be addressed to Jacquelyn B. Schriber, California School of Professional Psychology—Los Angeles, 2235 Beverly Boulevard, Los Angeles, California 90057.

Temporal Dimensions of Norms and Work Processes

Time and Work

Although there is some research and conceptualization about organizational culture and organizational norms, there is very little on the topic of time or norms about time—that is, about

time as the subject of research. The current U.S. concept of time can be characterized as homogeneous, objective, measurable, infinitely divisible, unidirectional, irreversible, reified, and linear (Byrne, 1982; Denhardt, 1986; Graham, 1981; Lauer, 1981; McGrath & Rotchford, 1983; Needham, 1966), although it may be occasionally viewed as cyclical (e.g., Lauer, 1981).

Two implications relevant to organizational behavior can be drawn from viewing time as was just described: Time is a scarce resource, and as a scarce resource, it must be managed. When time is viewed as a scarce, nonrenewable (though constantly available) resource (e.g., Becker, 1965; Fichman, 1984; Graham, 1981; Lauer, 1981; Moore, 1963), it attains both monetary and nonmonetary values (Becker, 1965; Calabresi & Cohen, 1968; Doob, 1971; Feldman & Hornik, 1981; Franklin, 1757/1973; Herzberg, 1971; McGrath & Rotchford, 1983; Moore, 1963; Yonge, 1973) that affect our behavior (Doob, 1971).

One process we use to control and manage the scarce resource of time is to divide it into segments that we allocate over a multitude of tasks or activities. Through this process we observe temporal patterns that provide a sense of orderliness (Zerubavel, 1981). The perceived orderliness reduces our feelings of uncertainty (Moore, 1963). Reducing uncertainty is an essential feature of organizational health and effectiveness (e.g., Bluedorn, 1986; Hirsch, 1980; Katz & Kahn, 1978; Perrow, 1970). Different organizations and individuals within them, however, manage time and reduce temporal uncertainty differently (e.g., Burke, 1986; Jacques, 1982; Puffer & Brakefield, 1986); these differences can be used to distinguish one organization or individual from another. For example, employees who take work home with them probably view their work differently from employees who keep strict time boundaries between work and leisure.

Organizations can and do manage time as they do other resources. Organizations may manage time both directly (e.g., through production scheduling), or indirectly (e.g., through investments in personnel strength, functional differentiation). Among the primary management processes that relate directly to the linear view of time are allocation, schedules and deadlines (including punctuality), pace, buffers, coordination, and synchronization. Less obvious but related temporal concepts include autonomy over the use of time, routine versus variety of tasks over time, and time boundaries (both between groups and between work time versus other types of time). These processes are implicitly understood in organizations, but rarely are made explicit and have not been studied. In addition, temporal dimensions related to the temporal values underlying a culture include temporal orientation (i.e., future, present, past), trade-offs between quality versus speed (related to work pace), and awareness of time use. A description of each of these temporal dimensions follows.

Temporal Dimensions

Allocation is the amount of time, whether planned or expended, devoted to an activity, regardless of when the amount occurs. It is based on a conventional standard of measurement (e.g., hours or minutes) and depends on an understanding of the

concept of duration—an understanding that seems evident even among primitive peoples (Doob, 1971).

Scheduling concerns location in the temporal realm (e.g., a 10:00 a.m. meeting) and gives organizations a framework for constructing temporal boundaries, just as yardsticks give builders a framework for constructing physical boundaries. Scheduling allows for the possibility of prediction and the resolution of temporal uncertainty (McGrath & Rotchford, 1983). The temporal boundaries provided by scheduling can be further defined in terms of sequence, deadlines, punctuality, pace, and buffers.

Sequencing refers to “actions follow[ing] one another in a prescribed order” (Moore, 1963, p. 8). Although scheduling is laying out a pattern of activities anchored to points in time within a specific time measurement system, sequencing is the ordering of activities over time within that system. This ordering may be inherent in the task (e.g., in the research process, data analysis precedes interpretation) or may be prescribed by the individual who controls the process. For example, some people take a shower before they get dressed in the morning, whereas others take a shower after they get undressed in the evening.

Deadlines are temporal start and stop points, and can be external or internal to the task, or both. Deadlines for single-activity tasks are based on temporal constraints external to the task. Deadlines for interdependent tasks are based both on temporal constraints external to the final task in the sequence and on temporal constraints internal to the task sequence itself (e.g., Gelles & Faulkner, 1978). A detailed discussion of deadlines as start and stop times of isolated and linked activities has been provided by McGrath and Rotchford (1983).

Punctuality is the degree of rigidity to which deadlines are adhered. Some deadlines require tasks to be completed on a certain day, others require completion by a certain hour of a specified day, and still others require completion by an identified minute of a particular day (e.g., fighter planes may be scheduled to take off from an aircraft carrier at precisely 14:07:00 hr).

Pace is the rate at which activities can be accomplished (i.e., the speed of activity or the number of activities that can be done within a given interval). Allocation, scheduling, and deadlines depend on pace. Each culture appears to have a pace that is considered appropriate for activity (Levine & Wolff, 1985).

Temporal buffers are unspecified amounts of time that are built into schedules to allow for the uncertainty in the estimated duration to accomplish a task or to allow for ease of scheduling on the basis of the imprecision of the time measurement system used. Buffers are evident in leads, lags, and delays (e.g., Cyert & March, 1963; Moore, 1963; Van Gigch, 1978; Zerubavel, 1981), and provide organizations with temporal elasticity, which in turn provides flexibility that may be necessary for survival (Moore, 1963; Van Gigch, 1978).

Autonomy over the use of time and time boundaries are also related to scheduling, although they are not direct characteristics of it. Both are secondary effects of scheduling, and reflect a more abstract level of the temporal environment. Autonomy is the amount of freedom the job holder has in setting schedules for the completion of his or her tasks over time.

Two types of temporal boundaries are relevant to work and organization culture. One type of boundary separates work

groups from one another; the other separates types of time from one another. As we noted earlier, the uses and meanings of time create group boundaries that indicate membership, values, and expected behavior patterns—the first type of boundary. Workers in one department may habitually work under tight schedules and experience pressure to be punctual, whereas another department may have a more temporally relaxed atmosphere.

The second type of temporal boundary separates work time from nonwork time for the individual. For some workers, these boundaries may be more permeable than they are for other workers. As Zerubavel (1981) noted, the private, nonwork time of low-status workers may be more institutionally protected than the private, nonwork time of professionals. Secretaries and clerks may typically “leave on time,” whereas their bosses “work late.”

The concepts of synchronization and coordination contain temporal dimensions that apply to work situations. For both, more than one task or activity is involved, and tasks may be performed by individuals or groups. Synchronization is managing the performance of more than one task simultaneously (e.g., to free a car when it is stuck in the snow, one must push *while* the car is in gear *and* the accelerator is depressed). In contrast, coordination may be thought of as managing the performance of more than one task in sequence (e.g., to produce a research report, questions must be identified, methods specified, data gathered and analyzed, interpretations made).

The 12 time-management concepts just described are consistent with the notion of time as linear, yet much work in organizations is not linear or unique. Few workers perform a given task only once during the course of their employment. The cyclical (recurring) nature of activities lies at the heart of ecological psychology (e.g., Wicker, 1979). A view of time that addresses time cycles adds depth to the temporal portrait of organizations (Schriber, 1985).

Routinization is one work concept that links linear and cyclical time. Yin (1979, p. 23) characterized routinization as “an organizational process that occurs over a long period of time in different organizations.” Yin emphasized the linear aspects; routinization is a linear temporal process abstracted in retrospect. At the same time, however, routinization is complete only when a “new” practice becomes fully integrated into the work flow. If a new practice were to be performed once or twice, it would not become routine. The process is analogous to the learning theory paradigm of shaping behavior through successive approximations (e.g., Millenson, 1967). Routinization implies repetition at appropriate times—it has cyclical characteristics both in the content of the activity and its placement in a schedule. The expectation is for smoother and more accurate performance of the behavior with each repetition, and perhaps for the eventual shortening of its duration or for an increase in pace. The opposite is variety (tasks rarely being repeated).

Finally, organization members’ awareness of their use of time on the job, the norms to which they adhere regarding trade-offs between the quality of work and the speed of work over time, and the temporal orientation of the organization as perceived by its members should reflect some underlying temporal values of an organization’s culture.

Study of Time At Work

If time is a key characteristic of organization culture, then it might be advisable to research time and culture using the same

methods. Presently, most organization culture research relies on qualitative approaches (e.g., Sathe, 1983; Schein, 1983; Smith & Simmons, 1983; Wilkins, 1983; Wilkins & Ouchi, 1983) or structured or semistructured interviews (e.g., Gregory, 1983; Marshall, 1982; Riley, 1983). These approaches are consistent with the concept’s roots in anthropology. There are, however, two shortcomings in these approaches. First, the number and identity of variables differ from one study to the next. Second, the number and identity of variables vary from one organization to the next, due to differences in the nature of the organizations under study. Thus, cross-organizational comparisons are difficult.

Because these shortcomings can be remedied by quantitative methods (which have their own shortcomings), we have chosen to use quantitative methods to study the temporal dimensions of work. As Jelinek and her associates (Jelinek, Smircich, & Hirsch, 1983, p. 331) pointed out, “Because the concept of culture in the study of organizations is not well developed, a range of approaches seems not only desirable but required.”

The purposes of this exploratory study are (a) to examine dimensions of time in organizations by examining temporally anchored norms—the perceptions of temporal rules and customs governing behavior in work organizations, and (b) to develop an instrument to measure the temporal dimensions of organization culture across different organizations, thus providing a way to compare organizations on this aspect of culture.

Method

Measuring Instrument

Most measurement techniques in the literature on the psychology of time or on time use in organizations did not lend themselves to the measurement of temporal norms or dimensions of organizations (e.g., Black, Bennett, & Wards, 1981; Calabresi & Cohen, 1968; Goodman, 1967; Guest, 1956; Jacques, 1982; Knapp & Garbutt, 1958; Mackay & Brown, 1970; Squyres, 1981; Stein, Sarbin, & Kulik, 1968; Yonge, 1973, 1974). Earlier work by Jackson (1966), however, used questionnaires to assess norms and roles in work organizations. Specifying contexts, situations, and behaviors with which the respondent could agree or disagree allowed for the assessment of norms. The instrument described ahead is based on that model.

A Time-At-Work questionnaire was designed to assess various possible temporal dimensions of work organizations. One section, Time Dimensions Scales, included 56, five-point Likert-type items that described time-related behaviors, situations, and customs concerning allocation of time, scheduling, temporal buffers, routine, autonomy, synchronization and coordination, temporal boundaries, work pace, time value, and future orientation. Respondents indicated their level of agreement, ranging from *strongly disagree* (1) to *strongly agree* (5), with each statement based on their perceptions of what was most typical within their work organization. Another section sought background information about individual respondents.

The questionnaire was self-administered; instructions for completion were located on the cover page. A letter explaining the purpose of the research accompanied each booklet. Respondents were promised anonymity. Pilot work indicated that time for completion ranged from 20 to 75 min.

Procedure

Organizational participation was sought from senior managers. Those who agreed to participate also served as research coordinators

for their organization. Participating organizations were located in California, Minnesota, North Carolina, Ohio, and Wisconsin, with some respondents from various organizations living and working in a variety of other states.

Each coordinator arranged the participation of at least two work groups from his or her organization; they were urged to make time available during the work day for questionnaire completion. Research coordinators were contacted within one week after the mailing of the questionnaires to verify receipt, and again during the second week as a reminder that prompt responses would be appreciated. The entire data collection effort spanned a 1.5-month period.

Subjects

To maximize the possibility of sampling a full range of differences among temporal norms, a sample was constructed to include variation along three dimensions: organization type, organization size, and work group type. First, to sample a variety of meaningful categories of work groups, participation from each of the five types described by Katz and Kahn (1978) was sought: production, supportive, maintenance, adaptive, and managerial. Second, organization size, which can be interpreted in a variety of ways (Porter, Lawler, & Hackman, 1975), was operationally defined as the organization's total number of employees. Third, respondents from organizations in all of three general operating function categories (manufacturing, service, and adaptive) were obtained to sample organization type.

Each questionnaire was coded for work group type and organizational membership, type, and size. A total of 529 completed questionnaires were returned from 51 work groups in 23 organizations (organizational identity could not be determined for 6 of the respondents). In all, there were 306 respondents from manufacturing organizations (10 firms), 173 from service organizations (8 firms), and 44 respondents from adaptive organizations (5 firms). Broken down another way, there were 162 respondents from production, 151 from supportive, 13 from maintenance, 96 from adaptive, and 101 from managerial work groups. Finally, if viewed from the organization size perspective, there were 399 respondents from large organizations (more than 500 workers—13 firms) and 124 respondents from small organizations (less than 500 workers—10 firms).

Respondents were predominantly white-collar, highly educated, well-paid, mid-career workers. Executives and managers composed 17% of the sample. Another 9.6% were supervisors or administrators; 8.9% were salespersons; 13.6% were clerical workers; 18.3% were business or other nontechnical professionals; 15.7% were engineers; 11.5% were technicians; and 5.1% were laborers.

The average education among the respondents was 15.4 years. Only 15.4% of the sample had not attended college; 23.3% had graduate school degrees. Earnings based on the previous year's job-related income were also high, with 69.1% of the respondents reporting earnings of \$20,000 to \$45,000 (the median income category was \$30,001–\$35,000). Only 16% of the sample reported earning \$20,000 or less; the remaining 14.9% earned over \$45,000. The average age was 38.1 years.

Most respondents reported working typical white-collar schedules. A Monday through Friday work week was reported by 82% of the sample. First shift was the norm for 98%. Less than 10% punched a clock, and 56% reported some travel as part of their job. Furthermore, 33% of the respondents supervised others. Among those with supervisory responsibilities, the median number of employees supervised was six; the mode, however, was one.

Results

Hypothesized Scales

In all, 56 Likert-type items were constructed to measure workers' perceptions of 15 separate temporal dimensions or

norms. The 15 hypothesized scales were as follows: the adequacy of the allocation of time for tasks, various aspects of scheduling (including punctuality, deadlines, and the sequencing of tasks), temporal buffers in both the workday and in planning, the synchronization and coordination of work with others through time, the perceived amount of routine in the job over time, temporal boundaries (both within the workplace and between work and nonwork time), the amount of autonomy over the use of time at work, the speed and pace of work, the awareness of using time as a resource, and the future orientation of the organization.

Time Dimensions Scales

To determine whether the 15 scales originally hypothesized were stable, and to determine if and how the reliability of each could be improved, factor analytic techniques were used. Because this was a field study in which a new instrument was being examined, principal components analysis was the method of choice. Principal components analysis explains the variance, rather than the covariance, in a set of data and is useful in reducing the number of variables based on their common variance (Kim & Mueller, 1978). In addition, principal components makes fewer assumptions about the data and does not tend to inflate the estimate of variance accounted for as compared with the common factor approach (Gorsuch, 1983; Green, 1978; Kim & Mueller, 1978; Tabachnick & Fidell, 1983). Common factor analysis would have been more appropriate if the instrument had previously been well explored and was known to be carefully balanced (i.e., if it were known that the scales on the instrument each had the same number of items and that the percentage of variance accounted for by each scale was approximately equal).

To be conservative, several analyses were performed, including principal components analysis, common factor analysis, and alpha factoring. The minimum eigenvalue measures associated with these tests suggested the extraction of 16 factors. In addition, items making weak contributions were systematically eliminated (see Schriber, 1985). The outcomes of these several analyses were similar, with items grouped into factors fairly consistently across approaches. A Procrustean rotation (Gorsuch, 1983; Green, 1978) of the principal components solution to the common factor solution as the target resulted in an average correlation between factors and components of .92. In addition, the amount of rotation required to move the principal components matrix to congruence with the common factor matrix was extremely small. Therefore, on the basis of both theoretical and empirical grounds, the principal components solution is the only one presented here. The results of this analysis, using a varimax rotation, are presented in Table 1.

In all, 16 principal components with eigenvalues greater than 1.00 were extracted and explained 59.0% of the variance. In all of the components, those items having negative loadings were worded in the opposite direction of those having positive loadings. The factors, therefore, were not bipolar. The means and standard deviations of acceptably reliable scales are displayed in Table 2.

Component 1 (Schedules & Deadlines, 7.4% of the variance) contained nine items having to do with the importance of meeting deadlines and the importance of staying on schedule and

Table 1
Correlations of Questionnaire Items in Principal Components Analysis (Varimax Rotation) With the Principal Component

Component and item	Eigenvalue	% variance	r	Component and item	Eigenvalue	% variance	r
1. Schedules & Deadlines	4.12	7.4		6. Awareness of Time Use (cont.)			
People here feel that deadlines don't really matter (R)			-.71	People here plan their time carefully (R)			-.55
Staying on schedule is important here			.67	People expect you to know how long it will take you to do something (R)			-.47
It is important to meet our deadlines			.67	7. Work Pace	2.11	3.8	
We don't pay much attention to schedules (R)			-.62	Working fast is not important here			.62
No one gets upset when you miss a deadline (R)			-.61	Most people can work at their own pace			.60
All of our work is tightly scheduled			.55	Most people can take breaks when they want to			.53
It is very important to be "on time" for everything			.53	It is easy to find time to plan something new			.48
People do most of their work under deadlines			.49	People are expected to work very fast (R)			-.44
People do things when they are ready, not on a schedule (R)			-.34	8. Autonomy of Time Use	2.05	3.7	
2. Punctuality	2.64	4.7		Around here, people like to talk about the "good old days"			.65
People get upset when you are late for work (R)			-.77	People here do not have the freedom to use their time the way they choose			.56
People don't care what time you arrive for work			.71	Most people here cannot set their own work schedules			.43
No one cares if you are late returning from a meal break			.70	People just expect to "kill time" on the job			.41
If people arrive an hour late for work, they will feel "rushed" all day (R)			-.51	People expect their work to be routine			.37
3. Future Orientation and Quality vs. Speed	2.45	4.4		9. Synchronization and Coordination of Work with Others Through Time	2.02	3.6	
This organization invests in the future			.70	To get the job done, it is important for each person to coordinate his/her work with others			.68
Planning for the future is important here			.69	People have to work together to get the job done			.68
Doing things right is better than doing things fast			.64	Teamwork is not very important here (R)			-.52
It is better to make a bad decision quickly, than a good decision slowly (R)			-.47	10. Routine vs. Variety	1.89	3.4	
4. Allocation of Time	2.29	4.1		People tend to do different things each day			.74
Schedules usually seem too tight for most big jobs/projects			.71	Our job duties seem to change from week to week			.62
We never seem to have enough time to get everything done			.64	Our jobs never seem to change very much (R)			-.52
Tasks usually take longer than planned			.63	People expect to finish their work by the end of each day (R)			-.34
5. Time Boundaries Between Work and Nonwork	2.28	4.1		11. Intraorganizational Time Boundaries	1.67	3.0	
People usually expect to take their work home with them			.73	Some departments work longer hours than others (R)			-.72
People expect to leave at the end of the day without worrying about their work			-.65	Everyone works about the same number of hours, no matter what jobs they hold			.67
People rarely get work-related calls during "off" hours (like nights and weekends) (R)			-.63	12. Time Buffer in Workday	1.56	2.8	
When people go on vacation, they are expected to tell their boss how to reach them			.56	Most people don't have time to take breaks during the day			.75
6. Awareness of Time Use	2.24	4.0		People could fit more into their workday if they had to (R)			-.52
Most people don't think about how they use their time			.66				
People here worry about using their time well (R)			-.63				

Table 1 (continued)

Component and item	Eigenvalue	% variance	r	Component and item	Eigenvalue	% variance	r
13. Sequencing of Tasks Through Time People can perform their tasks in any order and still get the job done (R) To get the job done, it is important to do tasks in a specific order	1.53	2.7	-.80 .69	15. Not named Around here, you can get your own work done, even when others don't	1.41	2.5	.79
14. Not named You can tell how important a person is by the hours he/she works Even major interruptions don't usually put us behind schedule	1.44	2.6	.75 .43	16. Not named Some people are free to set their own schedules, while others are not Around here, people are more concerned about their work for this month than for next year	1.35	2.4	.71 .38

Note. N = 473. (R) indicates that the scoring of the item is reversed.

being “on time.” Cronbach’s alpha for this nine-item scale was .80.

Component 2 (Punctuality, 4.7% of the variance) contained four items concerning norms about punctuality (e.g., the perceived effect of arriving late for work or late from meal breaks). Cronbach’s alpha for the four-item punctuality dimension was .74.

Component 3 (4.4% of the variance) combined two items intended to measure the future orientation of the organization with two items intended to measure the work pace. Because the two speed items were worded in opposite directions, the factor was not bipolar. However, the two speed items in this factor did not simply concern speed of work, but rather asked whether speed was more important than quality. Perceived future orientation of the organization (Future Orientation) and issues concerning trade-offs between speed and quality of work (Speed vs. Quality) seemed to be two separate concepts. Therefore, they were treated as two separate scales. The two items measuring Future Orientation had a Cronbach’s alpha of .74, and the two

items measuring the speed of work versus the quality of work had a Cronbach’s alpha of .63.

Component 4 (Allocation of Time, 4.1% of the variance) appeared to measure the adequacy of time allocation for tasks. The three items are concerned with too-tight schedules, not enough time to get things done, and tasks taking longer than planned. The reliability coefficient for this three-item scale was .59.

Component 5 (Time Boundaries Between Work and Non-work, 4.1% of the variance) emerged exactly as hypothesized and combined all four items originally intended to measure the strength of temporal boundaries between work and nonwork activities (between work life and personal life). The component measured the intrusion of job responsibilities into the worker’s personal time (Cronbach’s $\alpha = .65$).

Component 6 (Awareness of Time Use, 4.0% of the variance), as shown in Table 1, contained four items having to do with how much people plan and think about the use of their own time on the job (Cronbach’s $\alpha = .60$).

Table 2
Cronbach’s Alpha for Scales Derived From Principal Component Analysis

Principal component	Scale	n	Cronbach’s α	M	SD
1	Schedules & Deadlines	514	.80	3.9	0.7
2	Punctuality	523	.74	3.4	0.9
3	Future Orientation	524	.71	3.9	0.9
5	Time Boundaries Between Work and Nonwork	523	.65	3.6	0.9
3	Quality vs. Speed	490	.63	4.3	0.9
9	Synchronization and Coordination of Work with Others Through Time	522	.63	4.3	0.7
6	Awareness of Time Use	524	.60	3.4	0.7
7	Work Pace	522	.60	2.9	0.7
4	Allocation of Time	519	.59	3.6	0.9
13	Sequencing of Tasks Through Time	521	.58	3.4	1.1
11	Intraorganizational Time Boundaries	526	.55	3.7	1.0
8	Autonomy of Time Use	519	.54	2.4	0.7
10	Variety vs. Routine	521	.52	3.1	0.8

Note. A high score represents ■ high emphasis on Schedules & Deadlines, Punctuality, Future Orientation, Synchronization and Coordination of Work with Others Through Time, Awareness of Time Use, Sequencing of Tasks Through Time, and Variety vs. Routine, strong Time Boundaries Between Work and Nonwork, a stronger emphasis on Quality vs. Speed, and a low emphasis on a fast Work Pace, poor Allocation of Time, great disparity among hours worked by employees (Intraorganizational Time Boundaries), and least Autonomy of Time Use.

Component 7 (Work Pace, 3.8% of the variance) contained five items concerned with norms about the speed of work and pacing. Together, the five items have a Cronbach's alpha of .60.

Component 8 (Autonomy of Time Use, 3.7% of the variance) consisted of five items measuring the worker's autonomy over his or her own use of time while at work, time orientation, expectations about the value of using time at work, and the amount of routine in their jobs over time. When taken together, these items seemed to measure the degree to which respondents perceived autonomy over the use of their time. Cronbach's alpha for the five items was .54.

Component 9 (Synchronization and Coordination of Work with Others Through Time, 3.6% of the variance) combined three of the four items originally intended to measure the degree to which the synchronization and coordination of work with others over time is important (a temporally anchored concept of teamwork). Cronbach's alpha for these items was .63.

Component 10 (Routine vs. Variety, 3.4% of the variance) contained four items. It combined three of the items intended to measure perceived routine in the work- or task-mix over time with one concerning allocation. The reliability coefficient for the four items was .52.

Component 11 (Intraorganizational Time Boundaries, 3.0% of the variance) assessed the existence of different temporal boundaries within the organization (e.g., Do different departments work longer hours than others?) by combining two items originally intended to measure this aspect of the temporal dimensions of organizations (Cronbach's $\alpha = .55$).

Component 12 (Time Buffer in Workday, 2.8% of the variance) combined two of the items intended to measure this concept—the perceived amount of temporal buffer in the typical workday (Cronbach's $\alpha = .34$).

Component 13 (Sequencing of Tasks Through Time, 2.7% of the variance) emerged exactly as hypothesized and combined two items designed to assess the degree of temporal dependence of tasks within the respondent's job (Cronbach's $\alpha = .58$).

Components 14 and 16 (accounting for 2.6% and 2.4% of the variance, respectively), each of which combined two items, had no clear interpretation.

Component 15 (2.5% of the variance) contained only one item; the final communality of this item was quite high (.65), indicating that the single item was not a meaningful component (that is, it did not account for much unique variance).

The formation of scales using the Likert-type items was taken directly from the results of the principal components analysis. Both on the basis of the results of that analysis and the direction of scoring, coding of individual items was reversed where appropriate. Scale scores were obtained by adding the scores on each of the individual items within the scale and then dividing by the number of items in the scale. This procedure resulted in the 13 scales having a range of possible scores from 1.00 to 5.00.

Reliabilities of the Time Dimensions Scales

Three of the scales had very good levels of reliability, in the .70s and .80s range. Another seven scales showed mostly strong reliabilities—from the high .50s to the high .60s. Marginal reliabilities in the mid .50s were obtained for Autonomy of Time Use, Intraorganizational Time Boundaries, and Routine Versus Variety. Reliabilities for Components 12, 14, and 16 were ex-

tremely poor. In addition, the results of a scree test indicated that Components 12 through 16 could be considered trivial because a straight line through them indicated a major break in accounting for the variance in the factor solution (Gorsuch, 1983, pp. 166–167). Nonetheless, it appeared that Component 13 should be retained because (a) it consistently emerged in identical form over multiple analyses, (b) the principal components solution produced the scale exactly as hypothesized, and (c) the reliability of the scale was appreciably higher than that for Components 12, 14, 15, and 16. Table 2 presents summary information for each of the 13 usable scales, including the number of subjects on which each scale is based, the reliability coefficients (Cronbach's alpha), the means, and the standard deviations.

Discussion

The present study delineated 13 separate aspects of time in organizations, measured by 13 acceptably reliable scales having from two to nine items. Only 2 of the scales originally hypothesized were not recovered in the principal components analysis. Both dealt with the time buffer concept: time buffers as experienced in the workday and time buffers in the planning process.

On the basis of these results, it might be argued that the buffer concept—an unspecified amount of time built into a schedule to allow for uncertainty or imprecision in scheduling—is not applicable to time-related issues in organizations. Two more plausible explanations exist, however. One, the concept may not have been operationally defined sufficiently well to emerge as a stable result. Two, the respondents may not have been sufficiently aware of the presence or operation of buffers to report on them clearly. A principal component composed of two items intended to measure the time buffer in the workday did emerge. Unfortunately, reliability of this component scale was very low, and thus it was not included in the final set of scales. Nonetheless, this minor result indicates that the concept has some merit and that further work to measure it is warranted.

The Time Dimensions Scales from the Time-At-Work survey can be used in many ways. We suggest four: (a) comparisons within and across units, (b) comparison and examination of fit across levels, (c) further exploration for consistency in temporal norms, and (d) implications of temporal norms for other areas of research in organizational behavior.

First, the use of some or all of the scales can result in rich descriptions of the cultures of organizations or departments and can provide a basis for comparing organizations or subunits. Do similar departments (e.g., legal divisions, sales departments) across organizations have similar temporal norms? Are similar functional units across organizations more similar than two different functional units with the same organization?

A related issue concerns the temporal homogeneity versus heterogeneity within organizations. Do temporal norms apply throughout the organization or do they vary with level of hierarchy (i.e., Do executives adhere to different temporal norms than clerical workers? Is time allocated well in engineering but poorly in manufacturing?). Is autonomy in time use restricted to certain departments or occupations within the organization or are all units subject to the same temporal norms? Should they all be subject to the same norms? Which is associated with better performance, heterogeneity, or homogeneity of temporal

norms? How much latitude does the individual manager have in setting temporal norms or in changing them?

Second, the Time Dimensions Scales also could be used to study the fit between individuals and their organizational units, or between work groups and organizations. A person whose views of the importance of punctuality or schedules and deadlines differ from those of his or her colleagues might exhibit more withdrawal behaviors—low satisfaction, absenteeism, turnover—than other persons. The disparity between temporal norms of the organization and one's own temporal norms might be a useful predictor of withdrawal behavior or organizational or career success.

In addition, the scales can be used to examine the fit between ■ work group or department and the larger organizational unit. Are some work groups out of sync with respect to temporal norms, and what are the implications of their uniqueness? Such uniqueness might help consultants identify and locate pockets of innovators and entrepreneurs.

Third, the Time Dimensions Scales can be used to explore the temporal dynamics of organizations. How important (or unimportant) is time in defining an organization's culture? Is time a key aspect of culture or a minor aspect? Perhaps some temporal norms are more important than others to an organization's culture. Could such differences in the importance of various temporal norms across organizations or work groups be used to create typologies of work units?

Although the 13 temporal norms measured in this study are relatively independent of each other, there may be consistent or predictable patterns in certain situations. For example, a fast work pace may be associated with an emphasis on speed over quality, or routine over variety, or an acute awareness of time use. One might also expect that autonomy of time use would be associated with permeable time boundaries between work and nonwork.

A related issue concerns the association of patterns of temporal norms within occupations, work groups, or organizations. Do lawyers as a group, for example, show a consistent pattern of temporal norms that differ from managers, truck drivers, or assembly workers? At a different level of analysis, future research might show that service organizations display patterns of temporal norms different from those displayed by manufacturing firms.

Fourth, the Time Dimensions Scales can be used to explore the relations between temporal norms and many other topics in organizational behavior. For example, the scales could be used to test some common assumptions: Are manufacturing units that emphasize schedules and deadlines more productive, efficient, or profitable than other manufacturing units? Do different profiles on temporal dimensions that correspond with productivity, efficiency, or profitability relate to the technology or environment of the organization? Do organizations using future orientation survive longer than other organizations? Are they more or less profitable? Is poor allocation of time related to performance? Do organizational units exhibiting less autonomy in time use show less innovation and change than other organizations? Overall, an organization's temporal norms and time perspectives have implications for investments in long-range planning and research, for specific personnel policies, and for the fit between the organization and its external constituents.

The scale measuring temporal boundaries between work and

nonwork can be used to enrich our understanding of both domains. For example, perhaps employees who report distinct temporal boundaries between work and nonwork exhibit less commitment to their work, and are also more satisfied with their family activities or nonwork aspects of their lives. What are the nonwork consequences of working under strict norms of schedules and deadlines, punctuality, combined with poor allocation of time? What are the nonwork consequences of working under conditions of high autonomy in use of time and little emphasis on schedules and deadlines?

A planned change effort might also incorporate the use of the Time Dimensions Scales. For example, organizations that exhibit disparity in intraorganizational time boundaries and poor allocation of time may exhibit lower morale among employees. The Time Dimensions Scales could be used to uncover such organizational problems as may account for differences in withdrawal behavior from one organization to another. Similarly, the scales could be used to ferret out pockets of dissatisfaction that may lead to turnover within an organization. If schedules and deadlines are overemphasized in one unit, or one unit has especially little autonomy over the use of time in comparison to other units, these factors might account for employee withdrawal in those units. Similarly, combinations of temporal norms (e.g., poor allocation of time use, high degree of coordination and synchronization, fast pace, and emphasis on punctuality) may be associated with stress and burnout. Finally, combinations of temporal norms (e.g., Quality versus Speed, Sequencing of Tasks Through Time, Variety versus Routine, Awareness of Time Use) may be differentially associated with productivity depending on the work processes under consideration.

Overall, it is possible to measure people's experience of time at work. Future research should focus on validating the instrument and replicating the structure of the Time Dimensions Scales on different samples. Such additional work will be useful because research linking the temporal dimensions of work culture to other organizational concerns presents a broad array of options and promises to enrich our understanding of organizational behavior. How are the temporal dimensions of work associated with productivity, organizational conflict, effective job performance, or perceived stress? What implications do temporal dimensions hold for selection, placement, training, or work motivation? Future research could also focus on measuring and linking the nonlinear aspects of time as experienced within organizations with those dimensions discussed here. With this new way to look at organizations and work cultures, many possibilities exist.

References

- Becker, G. S. (1965). A theory of the allocation of time. *Economic Journal*, 75, 493-517.
- Black, W. A. M., Bennett, P., & Wards, A. R. (1981). Future events test: Equivalent forms and criminality. *Perceptual and Motor Skills*, 52, 277-278.
- Bluedorn, A. C. (1986, July). *Primary rhythms, information processing, and planning: Directions for a new temporal technology*. Paper presented at the meeting of the International Society for the Study of Time, Darlington Hall College of Arts, Totnes, Devon, England.
- Burke, C. G. (1986, August). *Stratified systems theory: A time based theory of organizations*. Paper presented at the meeting of the Academy of Management, Chicago.

- Byrne, N. T. (1982, September). *Time constructions in everyday life*. Paper presented at the meeting of the American Sociological Association, San Francisco.
- Calabresi, R., & Cohen, J. (1968). Personality and time attitudes. *Journal of Abnormal Psychology*, 73, 431-439.
- Cyert, R. M., & March, J. G. (1963). *A behavioral theory of the firm*. Englewood Cliffs, NJ: Prentice-Hall.
- Denhardt, R. B. (1986, August). *Time consciousness in organizational life*. Paper presented at the meeting of the Academy of Management, Chicago.
- Doob, L. W. (1971). *Patterning of time*. New Haven, CT: Yale University Press.
- Feldman, L. P., & Hornik, J. (1981). The use of time: An integrated conceptual model. *Journal of Consumer Research*, 7, 407-419.
- Fichman, M. (1984). A theoretical approach to understanding employee absence. In P. Goodman (Ed.), *Absenteeism: New approaches to understanding, measuring, and managing employee absence* (pp. 1-46). San Francisco: Jossey-Bass.
- Franklin, B. (1973). *Poor Richard's almanac*. New York: David McKay Co. (Original work published 1757)
- Gelles, R. J., & Faulkner, R. R. (1978). Time and television news work. *Sociological Quarterly*, 19, 89-102.
- Goodman, P. S. (1967). An empirical examination of Elliott Jacques' concept of time span. *Human Relations*, 20, 155-170.
- Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Graham, R. J. (1981). The role of perception of time in consumer research. *Journal of Consumer Research*, 7, 335-342.
- Green, P. E. (1978). *Analyzing multivariate data*. Hinsdale, IL: Dryden.
- Gregory, K. L. (1983). Native-view paradigms: Multiple cultures and culture conflicts in organizations. *Administrative Science Quarterly*, 28, 359-376.
- Guest, R. H. (1956). Of time and the foreman. *Personnel*, 32, 478-486.
- Herzberg, F. (1971). The motivation-hygiene theory. In D. S. Pugh (Ed.), *Organization theory* (pp. 324-344). New York: Penguin Books.
- Hirsch, P. M. (1980). Organizational effectiveness and the institutional environment. In D. Katz, R. L. Kahn, & J. S. Adams (Eds.), *The study of organizations* (pp. 175-192). San Francisco: Jossey-Bass.
- Jackson, J. (1966). A conceptual and measurement model for norms and roles. *Pacific Sociological Review*, 9, 35-47.
- Jacques, E. (1982). *The form of time*. New York: Crane Russak.
- Jelinek, M., Smircich, L., & Hirsch, P. (1983). Introduction: A code of many colors. *Administrative Science Quarterly*, 28, 454-467.
- Jones, G. R. (1983). Transaction costs, property rights, and organizational culture: An exchange perspective. *Administrative Science Quarterly*, 28, 454-467.
- Katz, D., & Kahn, R. L. (1978). *The social psychology of organizations* (2nd ed.). New York: Wiley.
- Kim, J., & Mueller, C. W. (1978). *Introduction to factor analysis*. Beverly Hills, CA: Sage.
- Knapp, R. H., & Garbutt, J. T. (1958). Time imagery and the achievement motive. *Journal of Personality*, 26, 426-434.
- Lauer, R. H. (1981). *Temporal man: The meaning and uses of social time*. New York: Praeger.
- Levine, R., & Wolff, E. (1985). Social time: The heartbeat of culture. *Psychology Today*, 19(3), 28, 30, 34-35.
- Mackay, C. K., & Brown, W. P. (1970). Metaphor preference vs. semantic ratings as measures of attitude toward time. *Journal of General Psychology*, 83, 207-212.
- Marshall, J. (1982). Organization culture: Elements in its portraiture and some implications for organization functioning. *Group & Organization Studies*, 7, 367-384.
- Martin, J., & Siehl, C. (1983). Organizational culture and counterculture: An uneasy symbiosis. *Organizational Dynamics*, 8, 52-64.
- McGrath, J. E., & Rotchford, N. L. (1983). Time and behavior in organizations. In B. Staw & L. Cummings (Eds.), *Research in organizational behavior* (Vol. 5, pp. 57-101). Greenwich, CT: JAI Press.
- Millenson, J. R. (1967). *Principles of behavioral analysis*. New York: Macmillan.
- Moore, W. E. (1963). *Man, time, and society*. New York: Wiley.
- Needham, J. (1966). Time and knowledge in China and the West. In J. T. Fraser (Ed.), *The voices of time* (pp. 92-135). New York: George Braziller.
- Perrow, C. B. (1970). *Organizational analysis: A sociological view*. Monterey, CA: Brooks/Cole.
- Porter, L. W., Lawler, E. E. III, & Hackman, J. R. (1975). *Behavior in organizations*. New York: McGraw-Hill.
- Puffer, S. M., & Brakefield, J. T. (1986, August). *Coping with time management problems: Dispositional and situational aspects of procrastination*. Paper presented at the meeting of the Academy of Management, Chicago.
- Riley, P. (1983). A structurationist account of political cultures. *Administrative Science Quarterly*, 28, 414-437.
- Sathe, V. (1983). Implications of corporate culture: A manager's guide to action. *Organizational Dynamics*, 12, 5-23.
- Schein, E. H. (1983). The role of the founder in creating organizational culture. *Organizational Dynamics*, 12, 13-28.
- Schriber, J. B. (1985). *An exploratory study of the temporal dimensions of work organizations*. Unpublished doctoral dissertation, The Claremont Graduate School, Claremont, CA.
- Smircich, L. (1983). Concepts of culture and organizational analysis. *Administrative Science Quarterly*, 28, 339-358.
- Smith, K. K., & Simmons, V. M. (1983). A Rumplestiltskin organization: Metaphors on metaphors in field research. *Administrative Science Quarterly*, 28, 377-392.
- Squires, E. M. (1981). Guidelines for use in scoring TAT stories for time span. *Perceptual and Motor Skills*, 52, 333-334.
- Stein, K. B., Sarbin, T. R., & Kulik, J. A. (1968). Future time perspective: Its relation to the socialization process and the delinquent role. *Journal of Consulting and Clinical Psychology*, 32, 257-264.
- Tabachnick, B. G., & Fidell, L. S. (1983). *Using multivariate statistics*. New York: Harper & Row.
- Van Gigch, J. P. (1978). *Applied general systems theory* (2nd ed.). New York: Harper & Row.
- Wicker, A. W. (1979). *An introduction to ecological psychology*. Monterey, CA: Brooks/Cole.
- Wilkins, A. L. (1983). The culture audit: A tool for understanding organizations. *Organizational Dynamics*, 8, 24-83.
- Wilkins, A. L., & Ouchi, W. G. (1983). Efficient cultures: Exploring the relationship between culture and organizational performance. *Administrative Science Quarterly*, 28, 468-481.
- Yin, R. K. (1979). *Changing urban bureaucracies*. Lexington, MA: Lexington.
- Yonge, G. D. (1973). Time experiences as measures of personality. *Measurement and Evaluation in Guidance*, 5, 475-482.
- Yonge, G. D. (1974). Dimensions of time experience. *Social Behavior and Personality*, 2, 119-124.
- Zerubavel, E. (1981). *Hidden rhythms: Schedules and calendars in social life*. Chicago: University of Chicago Press.

Received May 30, 1986

Revision received February 13, 1987

Accepted February 16, 1987 ■

Improving Group Performance by Training in Individual Problem Solving

Preston C. Bottger and Philip W. Yetton

Australian Graduate School of Management, University of New South Wales
Kensington, New South Wales, Australia

This laboratory study investigated the group performance effects of an intervention to improve individual members' use of task knowledge. Subjects were 169 managers (M age = 40 years; 90% men and 10% women) and 207 MBA students (M age = 28 years; 82% men and 18% women) working in 80 groups. The task was the *Moon Survival* problem. A bivariate, repeated measures analysis-of-variance statistical model was used. The intervention, based on suggestions by Maier (1970) and Janis and Mann (1977), upgraded both group resources (by 0.6 *SD*, $p < .05$) and group performance (by 0.6 *SD*, $p < .05$). The increment of group performance over average member scores was similar to that achieved by process and structural interventions reported in earlier studies. Supporting the proposition that group performance is strongly determined by member task ability, these results suggest that individual task training complements, if not supplements, the dominant process-training approach in a laboratory setting.

In small-group problem solving, the dominant research tradition has sought to identify process barriers to optimal resource use. Researchers have tried to overcome these limitations, either through training in group dynamics or by structural changes that eliminate face-to-face contact. Although structural interventions generally appear to upgrade performance, there is conflicting evidence about the impact of process consultation. The major reviews report weak effects of process intervention on performance in laboratory settings (Kaplan, 1979; Woodman & Sherwood, 1980). However, they do not reference a number of specific examples in which process training does substantially upgrade performance (e.g., Hall & Watson, 1970; Maier, 1970; Nemiroff & King, 1975).

In contrast to this interest in process consultation and structural intervention, studies of the effects of individual task training are rare in small-group research. This is perhaps not surprising given a paradigm that emphasizes process gains and losses and regards, at least implicitly, the relation between member ability and group performance as weak. In contrast, a recent sequence of studies has revived earlier interest in the role of member resources in determining group effectiveness (e.g., Einhorn, Hogarth, & Klempner, 1977; Laughlin, Kerr, Davis, Halff, & Marcinak, 1975; Yetton & Bottger, 1982, 1983). These studies suggest that manipulation of member expertise would have a substantial influence on subsequent group outcomes. The study reported here extends this research and examines the

effect on group performance of training group members to make more effective use of their existing knowledge.

A study examining the effect on group performance of training members to use their task knowledge effectively is novel in the small-group literature. Of course, in itself, this effect is neither surprising nor particularly noteworthy. Its significance, however, derives from its role in a larger set of research results (see Bottger, 1984; Bottger & Yetton, 1984; Yetton & Bottger, 1982, 1983). This research sequence questions the dominance of process explanations of group performance and argues that performance is also strongly determined by member task ability. The study we report here has the potential to strengthen this argument.

In order to facilitate the integration of this study with earlier articles in the series, the same experimental task is used: the *Moon Survival* problem. This choice has three advantages. First, as an extension to the earlier Bottger and Yetton studies, it represents a check on the internal validity of their *member task ability* model of group performance. If task training did not improve group performance, the internal validity of the model would be questioned. Second, this task has been used elsewhere to examine both process (Hall & Watson, 1970; Nemiroff & King, 1975) and structural (Gilmartin, 1974) interventions. These earlier studies provide performance data against which the present individual task training intervention can be evaluated. In addition, it provides the opportunity to investigate a threat to the construct validity of the process and structural interventions—the concept that both are confounded by individual task training. Although speculative, any such threat could be of major importance to the process and structural-intervention literatures.

Third, this task and its derivatives (e.g., *Desert Survival* and *Lost at Sea*) have been used extensively in management training to demonstrate that the superiority of groups over individuals is a function of creative error-correction processes that can be

This article is based on a paper presented at the Academy of Management Meetings, San Diego, California, August 1985, where it won the Outstanding Paper Award in the Management Education and Development division. The version presented here has benefited from comments and suggestions by two anonymous reviewers and by Irwin Goldstein.

Correspondence concerning this article should be addressed to Preston C. Bottger, AGSM, University of New South Wales, Kensington, NSW, 2033, Australia.

improved by process training (e.g., Hall & Watson, 1970). An alternative explanation and associated training would challenge that accepted interpretation and practice.

Research Context

Upgrading Individual Expertise

In their analysis of effective decision making, Janis and Mann (1977) focused on conflict and stress management. But, in contrast to traditional small-group research (e.g., Davis, 1973; Steiner, 1972), their analytical framework was within rather than between individuals. They argued that in making a decision an individual experiences internal conflict and stress. Furthermore, there is a risk of a low-quality decision if coping mechanisms are maladaptive. Such mechanisms include accepting or rejecting a solution without adequate search and evaluation, rejecting responsibility for the decision, avoiding corrective information, or panicking because of perceived lack of essential resources such as time and knowledge. In order to minimize such risks, they recommended training to increase information vigilance that promotes more effective use of the individual's resources.

A typical intervention, based on suggestions by Maier (1970) and Janis and Mann (1977), is as follows: First, participants are asked to reread the problem statement to ensure that they are solving the right problem. The instructor states that there is a high probability that a second solution to this problem would be superior to the first. It is then asserted that research has shown that four types of obstacles can limit problem-solving effectiveness: *hypervigilance*, *unconflicted adherence*, *unconflicted change*, and *defensive avoidance*. These are explained by the instructor. Hypervigilance—or panic—is discussed first in order to acknowledge participant anxiety and offer reassurance. The remaining topics are then discussed in turn by the instructor, who asks participants to consider the implications of these issues for their solutions. (The complete intervention is presented in the Method section.)

Within the group problem-solving framework, the effect of such training would be stated formally as follows:

Hypothesis 1. Group member resources are improved by an intervention that focuses on the effective use of the individual's task knowledge.

Bottger and Yetton (1984) argued that group performance is a positive function of member resources and the decision scheme for their use:

Group Performance

$$= \alpha_0 + \alpha_1 \text{Resources} + \alpha_2 \text{Decision Scheme}, \quad (1)$$

where Resources are measured as the mean of the best two member solutions at the item level (an "expert supported" solution; cf. Laughlin et al., 1975) and Decision Scheme is the deviation between the group solution and that defined by the mean of the best two members (cf. Davis, 1973). A full rationale for the model and measures is given elsewhere (Bottger & Yetton, 1984). This model suggests that an increase in member resources through individual training would result in increased group performance. We can state this formally as follows:

Hypothesis 2. The problem-solving performance of groups

whose members are trained with an individual-based task intervention is superior to groups with untrained members.

Method

Experimental Task

The Moon Survival task requires subjects to imagine themselves crash-landed on the moon 200 mi. from base. Fifteen pieces of equipment are available for use and are to be ranked in order of declining contribution to survival on the walk to safety. Performance is a simple inverse function of the unit-weighted sum of the absolute differences between the ranks assigned and the ranks preferred by the Crew Equipment Research Unit at the National Aeronautics and Space Administration. Thus, a low score indicates high performance.

This exercise is more difficult than some tasks used in group research. For example, Laughlin et al. (1975) use a word-synonym problem with demonstrably correct answers. Solutions to Moon Survival require knowledge of the moon environment and of navigation and survival techniques, and the typical subject probably knows less about such topics than about vocabulary. However, solving both types of problem essentially requires knowledge of relevant facts. Therefore, differences in the demonstrable correctness of solutions between Moon Survival and a vocabulary test seem largely an issue of degree rather than kind.

Similar to other research using this task (e.g., Gilmartin, 1974; Hall & Watson, 1970), most subjects appeared highly involved in the problem-solving process. Postexperimental subject reports confirmed that the exercise was seen as a test of individual task knowledge and of small-group discussion skills. The task seems to be analogous to managerial problems in which knowledge of the problem space varies across group members, and in which, initially, members are unaware of their colleagues' relative task expertise.

Subjects

Subjects were a total of 376 managers and masters of business administration (MBA) students working in 80 groups. Of these, 249 individuals in 53 groups from the existing data base (Bottger, 1984; Bottger & Yetton, 1984; Yetton & Bottger, 1982, 1983) were in the *untrained* condition. In addition, 127 individuals in 27 groups received the intervention in the *trained* condition. There were no mixed manager and MBA groups. Sample composition and subject details are reported in Table 1. (Pre- and postintervention group resources, decision schemes, and performance scores are shown in Table 2, and mean individual scores are reported in Table 3.)

As discussed in detail elsewhere (Bottger, 1984; Yetton & Bottger, 1983), there were trivial nonsignificant differences in individual and group scores across groups with different membership characteristics. Also, correlation matrices for the study variables were very similar for the manager and student subsamples. As such, the data were combined for analysis. Validity threats owing to different composition profiles between the untrained and trained groups are considered later. Finally, there is no evidence that Australian subjects differ from the U.S. subjects of the previous research studies that used the Moon Survival problem (Bottger, 1984).

Procedure

Assignment of subjects. The difference in numbers between the two conditions is unusual for a laboratory study in which allocation of subjects is controllable and, as such, requires explanation. As noted, this study is one of a series in which we examine group performance. We did not begin our group research with the present study specifically in mind; the idea for it developed as our work progressed. Thus, when we

Table 1
Sample Composition and Subject Details

Sample characteristic	Managers	MBA students
<i>N</i> subjects	169	207
<i>N</i> groups ^a	36	44
Age ^b		
<i>M</i>	40.0	28.0
<i>SD</i>	5.9	5.3
Work experience ^b		
<i>M</i>	20.0	4.0
<i>SD</i>	7.6	3.8
Sex ^c		
% men	90	82
% women	10	18
Nationality		
% Australian	100	75
% Asian	0	25

Note. MBA = Masters of Business Administration.

^a There were 30 groups with 4 members, 44 with 5 members, and 6 with 6 members.

^b Age and work experience in years; 20% of the MBA students had not held a full-time job.

^c Most groups had 1 female member.

began testing the intervention, we already had 53 untrained groups that could be used for baseline data. The use of these data is explained later.

Experimental conditions. The procedures, with the exception of the intervention, were identical for both the trained and untrained conditions. The task was represented to all subjects as an exercise in individual and group problem solving. It was one of a sequence of activities consisting of courses in management development for the managers and managerial psychology for the students. In the untrained condition, subjects first solved the problem individually and then were assigned randomly to groups. We allowed 10 min for the individual phase and 30 min for the group discussion. These times are typical where the Moon Survival task is used for research and training (e.g., Hall & Watson, 1970). Groups were given identical written and verbal instructions in order to develop a team solution in whatever way they thought appropriate. Leaders were assigned randomly and instructed to coordinate the discussion, keep time, and enter the group solution on the answer sheet.

The trained condition included an additional step. After subjects first solved the problem individually, the instructor (one of the authors) read a statement explaining threats to decision adequacy. Subjects were asked to consider these threats while listening to the instructor's statement, to rethink their initial solution against the possibility of such threats, and to make notes where appropriate. After this intervention, additional time was allowed for resolving the problem. We allocated 10 min for the explanation phase and 10 min for the second solution. Thus, the trained groups do not solve the problem *as a group* without training. Therefore, for trained groups, actual preintervention performance and decision-scheme data (see Equation 1) are not available and must be estimated. This is discussed below.

The Intervention: Threats to Vigilant Information Processing

Two instructors, the authors, administered the intervention, each training approximately half of the 27 groups in the experimental condition. These were treated in 5 subgroups of 5 or 6 groups each. In order to maintain consistency in the intervention across the 5 subgroups, we read the same instructions and closely conformed to the following format:

Are you solving the right problem? First, reread the problem statement and check whether you are solving the problem as stated. Check whether you have misread the instructions or made unwarranted assumptions.

Research on the problem-solving process suggests that there are four common ways of getting a poor solution to a problem like this. I am now going to discuss these in turn and ask you to reconsider your first solution to this problem in the light of what I say. These ways of achieving poor solutions are:

- **Hypervigilance.** This pattern is marked by frantic searches for quick solutions. You might switch from one line of thought to another without learning from either. It might be difficult to think clearly. The state of hypervigilance (e.g., the experience of memory loss, high levels of emotional excitement, and repetitive, inefficient thinking) is not uncommon. Often, a last chance solution is accepted because it offers relief from anxiety and uncertainty. This solution is frequently a poor one, because it does not benefit from proper analysis and evaluation.

If you find yourself behaving like this, keep these points in mind:

1. The time available to do this task, to attempt to improve your original solution, is adequate for a thorough reexamination of the facts. After all, you will have a total of 20 min—that is, 10 min more than was available for the first attempt. The only real issue is how to allocate this time.

2. Given your level of education and information base, it is highly probable that you can construct a very good solution.

3. We are now going to discuss the techniques for improving your solution. You need only apply them to get a good result.

- **Unconflicted adherence.** This response amounts to your remaining anchored on your present ranks. A common problem for decision makers is that they stick with the first idea that comes into their heads, without further evaluation of its consequences. This can be dangerous because information about potential losses or risks are ignored or not sought.

If unconflicted adherence is likely to be your behavior, try the following:

1. Search your knowledge base again and check if you know more about the problem environment. What elements of your solution are absolutely necessary for achieving a successful result? What difficulties are likely to be encountered in implementation? What can you afford to take a risk on?

2. Avoid thinking "it does not matter about this particular part of the solution" until you have actually assessed the risks in ignoring it.

3. Try turning the problem, or its solution, on its head. Can the reverse tell you anything about the real situation?

4. Make sure that you are not trying to solve the problem with an inappropriate mental set. Ensure that you are not importing preconceptions from previous problem situations. How is this decision space unique? How is it different from your earlier experience?

- **Unconflicted change.** You might change your mind uncritically and accept the first new idea that comes along. If you find yourself doing this, ask these questions:

1. Are there further risks of losses or poor results deriving from this new solution?

2. Was my first solution so bad?

3. What are the likely differences, in terms of good and bad outcomes, among this new solution, my original solution, and the next plausible alternative?

- **Defensive avoidance.** There are three forms of this response: procrastination, disowning responsibility, and selective inattention to corrective information.

1. Procrastination. This is a likely response if you cannot imme-

diately cope with your uncertainty about the relative values of alternative solutions. There is nothing intrinsically wrong with taking your time over a judgment, but on this problem, as in real life, there are certain time limits. Spread your available time over the decision space in a way to maximize reward for effort.

2. Disowning responsibility. In this problem, disowning responsibility might take the form of saying "Oh well, this is only a classroom problem. No point in trying to do well. Who cares?" The answer is that willingness to try out new behaviours on simple problems like this is the basis for improvement on a larger scale.

3. Selective inattention to corrective information. If you feel uncertain about your solution, but want to be finished with the problem, you might tend to ignore some of your nagging doubts. Doubts can be useful in decision making. Use them to test the robustness of your judgment, set them out side by side—on a piece of paper if necessary—in a balance sheet. Does a particular pro argument outweigh a particular con argument? Does one con argument cancel several pro arguments?

Now, finalize your second solution to the problem.

Estimation of Baseline Control Data

In the later analysis, untrained groups are not compared directly with trained groups to examine the intervention effect. Rather, data on the 57 untrained groups are used in the model of group performance (Equation 1) to estimate coefficients for group resources (the expert-supported solution) and decision scheme (the deviation between the group solution and that defined by the expert-supported solution). Preintervention group resource data and decision-scheme data for the trained groups are then substituted in the resulting equation to estimate the group performance that could be expected had the intervention not been applied. Then, in the analysis of intervention effects, each group acts as its own control. Postintervention group resources and performance are compared with actual preintervention resources and estimated preintervention performance, respectively. This procedure helps overcome one possible threat to the internal validity of the findings—that the intervention effect might be explained by regression to the mean caused by the relative inferiority of preintervention trained-group resources (see Table 2).

The estimation of preintervention performance for trained groups involves several assumptions. First, the model specified in Equation 1 is assumed to be an adequate representation of group performance (regression results are presented later). This assumption includes a decision to perform the analysis at the item level of the Moon Survival problem. That is, group performance for each of the 15 items in each group is regressed on the respective item's group resource and decision scheme. This procedure is consistent with arguments that item-level modeling is more appropriate than task-level analysis for multipart problems, given within-group variability across parts in performance, resources, and decision schemes (see Bottger & Yetton, 1984; Laughlin et al., 1975). Although the modeling and estimation are undertaken at the item level, all results are reported at the task level (by multiplying each number by 15) for purposes of comparison with earlier studies.

Second, we assume that trained groups, had they not been trained, would use their resources in the same way as the untrained groups. This might introduce a threat to internal validity, given the difference in non-trained resources between untrained and trained groups (see Tables 2 & 3). However, Bottger and Yetton (1984) report a weak relation between average resources and decision scheme, suggesting that this second assumption is robust. Third, we assume that the decision schemes used by preintervention trained groups equal their postintervention decision schemes. That is, we assume that the task intervention does not affect decision scheme. In the Discussion, we address a possible internal-validity threat posed by this assumption.

Results

As explained earlier, expected trained group performance, had the intervention not been applied, is calculated by (a) estimating the coefficients of Equation 1, using the data on the 53 untrained groups at the item level, (b) substituting the trained groups' preintervention resources and postintervention decision-schemes data in the resulting equation, and (c) scaling up the results by a factor of 15 for performance at the task level. That is,

Group Performance

$$= 15(0.25 + 0.60 \text{ Resources} + 0.81 \text{ Decision Scheme}). \quad (2)$$

The regression details for this equation are as follows: $R^2 = 0.69$, $F(2, 792) = 866.3$, $p < .05$; $t(792; \text{Resources}) = 19.1$, $p < .05$; $t(792; \text{Decision Scheme}) = 33.7$, $p < 0.5$.

The hypotheses are tested using a bivariate (group resources and group performance) repeated measures (pre- and postintervention) ANOVA. Each group acts as its own control. Table 2 reports means, standard deviations, and the pre- and postintervention group resources, decision scheme, and performance scores.

Hypothesis 1 is supported: The intervention improves "expert-supported" group resources from 20.6 to 16.8 error points, $F(1, 26) = 19.6$, $p < .05$ (repeated measures ANOVA). The data in Table 2 also support Hypothesis 2: Trained group performance is superior to that predicted using preintervention group resources (26.6 vs. 31.5 error points), $F(1, 26) = 21.1$, $p < .05$ (repeated measures ANOVA).

Discussion

The individual-ability intervention improves group resources and performance by more than half of a standard deviation. These improvements in individual performance replicate Maier's (1933, 1970) findings that, for both individuals and groups, second and subsequent attempts at problem solving, with guidance in the effective use of information, are likely to produce better solutions than first tries. Simply asking subjects to reconsider initial solutions does not appear to help as much as an intervention that imposes structure and direction on the thinking process (Gilmartin, 1974; Maier, 1970; Van den Ven, 1974).

Earlier, we noted a potential threat to the internal validity of these findings through differences in untrained and trained group resources. We argued that using each group as its own control reduced this threat. This procedure introduced another threat to internal and statistical validity by assuming that preintervention trained-group decision schemes equaled their respective postintervention decision schemes. This is a conservative assumption, however, which decreases rather than increases the probability of finding an intervention effect on group performance.

Elsewhere, Bottger and Yetton (1984) reported that, whereas it is trivially associated with the level of group resources (measured as the mean of the best two members scores or the average member score), decision scheme is a function of resource composition. Here, composition refers to the incidence of decision

Table 2
Means and Standard Deviations of Resources, Decision Scheme, and Group Performance for Trained and Untrained Groups

Group	Pretraining			Posttraining		
	Resources	Decision scheme	Group performance	Resources	Decision scheme	Group performance
Untrained						
<i>M</i>	15.9	23.0	29.1			
<i>SD</i>	5.2	6.9	8.3			
Trained						
<i>M</i>	20.6	22.8 ^a	31.5 ^b	16.8	22.8	26.4
<i>SD</i>	6.2	6.4	8.6	5.5	6.4	7.6

^a Assumed equal to posttraining value.
^b Estimated from Equation 1 (see Research Context section).

items for which there are majority-correct, pair-correct, no-plurality, and majority-incorrect member solutions. The effectiveness of the decision scheme declines in the order of the conditions just listed. This has two implications. First, knowledge training increases the proportion of majority and pair corrects and therefore improves the group decision scheme. As such, the pretrained-group performance scores calculated above are overestimated, with unknown bias. The actual effect of the intervention is therefore stronger than reported. Second, as discussed later, knowledge training for individuals might not be simply complementary to structure of process interventions; instead, it could be a substitute.

These findings contribute to the coherent picture of results from our earlier studies. Throughout these studies, the pattern of findings points to a strong effect of member ability on group effectiveness. Yetton and Bottger (1982) showed that groups can identify relative member expertise and use it in problem solving. Yetton and Bottger (1983) found that group composition based on task skill has a greater impact on performance than team size or decision scheme. Bottger and Yetton (1984) showed

that group performance strongly depends on the presence of expertise. Bottger (1984) found that when participation levels are distributed proportionally to expertise, groups use rational decision schemes and attain high performance.

The study we report here provides a check on the internal validity of these results. If a variable other than ability actually accounted for the findings mentioned above, an intervention to upgrade member ability would be unlikely to improve group effectiveness. These results, confirming a positive effect of such an intervention on performance, are entirely consistent with and reinforce the earlier findings that group products are highly dependent on member task contributions.

The results of individual training in this study can be compared with the effects of process and structural interventions in previous research. Data on expert resources and decision schemes are not available from those earlier studies. However, a comparison can be made of improvements of group performance over average member ability across all studies. Table 3 shows Moon Survival performance scores for groups in four categories: untrained, individual training, group-process training,

Table 3
Moon Survival Problem: Error Scores for Individuals and Groups Under Conditions of No Training and of Task, Process, and Structural Interventions

Group type	Average member score		Average group score	Improvement of group scores over pretraining member scores
	Pretraining	Posttraining		
No intervention				
<i>M</i>	40.4		29.1	11.3
<i>SD</i>	11.6		8.3	
Individual task intervention				
<i>M</i>	44.3	41.6	26.4	17.9
<i>SD</i>	12.1	11.8	7.6	
Group process intervention (Hall & Watson, 1970) ^a	45.1 ^a		25.6	19.5
Group process intervention (Nemiroff & King, 1975) ^a	41.8		26.5	15.3
Group structure intervention—SPAN (Gilmartin, 1974) ^a	44.6		24.6	20.0

Note. The no-intervention and individual task intervention rows report results from the study reported in this article.
^a Standard deviations not reported in these studies.

and a structural intervention known as SPAN. Although untrained teams outperform their average member by 11.3 error points, the other groups show gains of 17.9, 17.4 (average over two studies) and 20.0, respectively. That is, individual training, group-process training, and a structural intervention all produce similar improvements in performance.

This raises two issues. One is the relative efficiencies of these apparently different training methods. The other concerns the extent of overlap among them. In order to set the context for an examination of these two issues, we will now briefly review the critical elements in process and structural interventions.

Process is defined here as the system of intermember verbal and nonverbal exchanges that influence group outcomes (e.g., Dunphy, 1972; Hackman & Morris, 1975; Van den Ven, 1974). In this literature, conflict management is regarded as an important dimension of group process. For example, Hall and Watson (1970) developed a process intervention to manage intragroup conflict and investigate its effects on the Moon Survival problem. They argued that positive conflict management involves the examination of competing knowledge bases, exploration of alternatives, and the willingness of participants to argue for their point of view. That is, there is an emphasis on expertise, logical argument, and explanation. By contrast, negative conflict management, which should be discouraged, includes "I win, you lose" dominance gains, voting or coin tossing to resolve opinion differences, and the reluctance of some participants to argue for their points of view.

In comparison, *structural interventions* are based on the implicit or explicit assumption that process losses through poor group dynamics are not only substantial, but difficult if not impossible to correct. Examples of structural interventions are the Delphi technique (Dalkey, 1969), the nominal group procedure (Van den Ven, 1974), and the Social Participatory Allocative Network (SPAN) technique (McKinnon & Cockrum, 1973). The SPAN technique is described as the example here, because it was used by Gilmartin (1974) to compare the relative effectiveness of process and structural interventions on group performance on the Moon Survival exercise. His intervention is designed to take advantage of group resources and reduce or eliminate negative process effects.

SPAN requires individuals to judge both the relative quality of alternative solutions to a problem and the relative task expertise of group comembers. Each member begins with a parcel of votes. These are allocated to solutions and members (or to both) in proportion to the individual's judgment of solution quality or member expertise. A computer program redistributes to solutions the votes assigned to a member in proportion to that individual's initial distribution of votes. The result is a set of weights for solution options reflecting member judgments both of particular solutions and other participants' capacity to make judgments about those solutions.

Issues of relative efficiencies and overlap among different interventions were raised earlier. These brief descriptions of process and structural interventions suggest that SPAN is expensive relative to individual and process techniques. It requires access to the SPAN program and a computer. Furthermore, each problem presumably requires reprogramming. In contrast, knowledge and process training are both simply administered and are generalized across problems. That is, accessing SPAN incurs

recurrent costs, whereas group members could quickly become trained in the other two techniques.

In addition, both the descriptions above, and the similar performance effects of the three interventions, suggest that they are not independent of each other. Specifically, there are similarities between both process and structural interventions and the rational nonunit-weight decision strategy proposed by Einhorn et al. (1977) and extended by Yetton and Bottger (1982, 1983). Although nonunit weighting occurs in the context of interacting groups and SPAN controls such interactions, both are based on the recognition of differential task ability and the differential weighting of such expertise. SPAN requires that group members can make accurate judgments about relative expertise to compute weightings. Bottger and Yetton (1984) found that effective interacting groups act as if they use a nonunit-weight decision scheme, and that performance is a positive function of conformity with it.

It is also possible to reinterpret the process interventions, at least partially, as based on use of member abilities. Maier's (1970) interventions are derived from individual-training techniques and Hall and Watson's (1970) interventions, also used by Nemiroff and King (1975), direct members' attention to expertise and its effective use. Furthermore, in following such training, members are likely to reassess their own knowledge in much the same way as advocated by Janis and Mann (1977), although perhaps less effectively. Similarly, in applying SPAN, members are asked to evaluate each other's expertise. In both cases, there is a possible confounding with a knowledge-based intervention.

Finally, it is interesting to speculate how these laboratory results could generalize to organizational settings. It might be argued that the political nature of key organizational decisions would minimize, if not negate, the usefulness of a knowledge intervention. Demands of competing self-interests and the complexities of multiobjective, uncertain decision spaces (or both) could greatly reduce the positive effect of such an intervention. We can only agree with such an argument. As such, the scope for application of the knowledge intervention might be limited to those situations in which process interventions would work—that is, conditions under which the participants' goodwill and orientation to effectiveness are well developed.

Summary and Conclusions

Studying individuals and groups working on the Moon Survival exercise, we found that task training for individuals—a technique neglected in the current literature—improves group problem-solving performance. By training individuals in effective search and evaluation routines before they assemble in groups, member use of task knowledge is upgraded and team achievement is enhanced. This approach to group problem solving is different from the currently dominant orientation in the literature, which prescribes process consultation to enhance group effectiveness. However, little evidence supports such a view (Kaplan, 1979). As argued above, such interventions appear to work when their focus overlaps with consultations centering on the proper use of task knowledge. Group problem-solving performance might be improved, at least as efficiently

as if not more effectively than process consultation, by an intervention that improves member task contributions.

References

- Bottger, P. C. (1984). Expertise and air-time as bases of actual and perceived influence in problem solving groups. *Journal of Applied Psychology*, 69, 214-221.
- Bottger, P. C., & Yetton, P. W. (1984). *Group problem solving: Roles of resources, strategy and creativity* (AGSM Working Paper Series 84-20). Sydney, Australia: University of New South Wales, Australian Graduate School of Management.
- Dalkey, N. (1969). Analyses from a group opinion study. *Futures*, 1, 541-551.
- Davis, J. H. (1973). Group decision and social interaction: A theory of social decision schemes. *Psychological Review*, 80, 97-125.
- Dunphy, D. C. (1972). *The primary group: A handbook for analysis and field research*. New York: Appleton.
- Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgement. *Psychological Bulletin*, 84, 158-172.
- Gilmartin, K. M. (1974). The relative effectiveness of SPAN and laboratory training in upgrading decision making (Doctoral dissertation, University of Arizona, 1974). *Dissertation Abstracts International*, 35, 4624B.
- Hackman, J. R., & Morris, C. G. (1975). Group tasks, group interaction process and group performance effectiveness: A review and partial integration. In L. Berkowitz (Ed.) *Advances in Experimental Social Psychology* (Vol. 8, pp. 47-99). New York: Academic Press.
- Hall, J., & Watson, W. H. (1970). The effects of a normative intervention on group decision-making performance. *Human Relations*, 23, 299-317.
- Janis, I. L., & Mann, L. (1977). *Decision making*. New York: Free Press.
- Kaplan, R. E. (1979). The conspicuous absence of evidence that process consultation enhances task performance. *Journal of Applied Behavioural Science*, 15, 346-360.
- Laughlin, P. L., Kerr, N. L., Davis, J. H., Halff, H. M., & Marcinak, K. A. (1975). Group size, member ability, and social decision schemes on an intellectual task. *Journal of Personality and Social Psychology*, 32, 522-535.
- Maier, N. R. F. (1933). An aspect of human reasoning. *British Journal of Psychology*, 24, 144-155.
- Maier, N. R. F. (1970). *Problem solving and creativity in individuals and groups*. Belmont, CA: Brooks/Cole.
- McKinnon, W. J., & Cockrum, D. L. (1973). SPAN 2: Modification of SPAN Program for Synthesising Group Decisions. *Behavioural Science*, 18, 78-79.
- Nemiroff, P. M., & King, D. C. (1975). Group decision making: Performance as influenced by consensus and self orientation. *Human Relations*, 28, 1-21.
- Steiner, I. D. (1972). *Group process and productivity*. New York: Academic Press.
- Van den Ven, A. H. (1974). *Group decision making and effectiveness*. Kent, OH: Kent State University Press, Comparative Administration Research Institute of The Center for Business and Economic Research, Graduate School of Business Administration.
- Woodman, R. W., & Sherwood, J. J. (1980). The role of team development in organizational effectiveness: A critical review. *Psychological Bulletin*, 88, 166-186.
- Yetton, P. W., & Bottger, P. C. (1982). Individual versus group problem solving: An empirical test of a best member strategy. *Organisational Behaviour and Human Performance*, 29, 307-321.
- Yetton, P. W., & Bottger, P. C. (1983). Relationships amongst group size, member ability, decision schemes and performance. *Organisational Behaviour and Human Performance*, 32, 145-159.

Received January 2, 1986

Revision received April 17, 1987

Accepted November 22, 1986 ■

Comparative Analysis of Goal-Setting Strategies Across Cultures

Miriam Erez

Technion-Israeli Institute of Technology, Haifa, Israel

P. Christopher Earley

Department of Management and Policy, University of Arizona

Only a few studies that have examined the effects of participation on an individual's goal acceptance and performance have been conducted within a cross-cultural context. In the present study, we tested for the contingency between the effectiveness of goal-setting strategies and cultural values. We examined three goal-setting strategies within three different cultural groups—assigned goals, goals participatively set by a group representative and the experimenter, and goals participatively set by a group. The three cultural groups studied were U.S. students ($n = 60$), individualistic and having a high power distance; Israeli students from urban areas ($n = 60$), collectivistic and having a low power distance; and Israeli students from kibbutzim ($n = 60$), highly collectivistic and having a low power distance. Results indicated that participative strategies led to higher levels of goal acceptance and performance than the assigned strategy. Culture did not moderate the effect of goal-setting strategies on goal acceptance, but it appeared to moderate the effect of strategy on performance for extremely difficult goals.

Few studies that have examined the effects of participation in goal setting on an individual's productivity have been conducted within a cross-cultural context (Erez, 1986). Most of the studies concerning participation have been conducted at the societal level, focusing on participative and collectivistic values (Hofstede, 1980; Ronen & Shenkar, 1985) or on organizational level practices and their effects on work attitudes (Haire, Ghiselli, & Porter, 1966; Heller & Wilpert, 1981; Tannenbaum, Kavcic, Rosner, Vianello, & Wieser, 1974). Extending the previous lines of inquiry concerning worker participation, we examined the role of cultural values in relation to management practices at the microlevel of individual behavior.

The importance of cultural values in determining an individual's reactions in the workplace has been suggested by several researchers. Wilpert (1984) developed a conceptual framework in which participative management practices are examined within the context of cultural values; he emphasized that the levels of societal values and managerial practices correspond to one another. Investigations on participation conducted by European researchers (IDE, 1981), as well as others (Maurice, Sorge, & Warner, 1980), suggested that the strongest predictors of attitudes toward participation and of actual involvement are institutionalized norms and experience with a participative approach (French, Kay, & Meyer, 1966). According to Locke and Schweiger (1979) employee participation in Europe, unlike that in the United States, is institutionalized by law and is anchored in the political system that advocates socialistic and egalitarian values.

A contingency approach relating cultural values to managerial practices suggests that the use, as well as the effectiveness,

of participation is influenced by the norms prevalent in a society. For instance, Erez (1986) examined the influence of cultural differences in individualistic versus collectivistic values on the effects of participation. Participative and nonparticipative strategies for setting performance goals were used to motivate employees in three organizational subcultures—kibbutz (commune), Histadrut (trade union), and private sector in Israel—known to differ in collectivistic values. The results demonstrated that the nonparticipative strategy was the most effective in the private sector, participation by representative was the most effective in the trade union sector, and group participation was the most effective in the kibbutz, the most collectivistic sector.

The contingency approach may explain why participation appears to be more beneficial in certain cultures than in others. For instance, the impact of cultural values on the perceived legitimacy of goal-setting strategies and consequent performance was discussed by French, Israel, and Ås (1960). In an attempt to explain their failure to replicate the findings of Coch and French (1948), French et al., (1960) argued that the Norwegian workers in their study were unionized, and perceived participation by individual workers as being inconsistent with the philosophy of participation through union representatives. The workers viewed the union representative as the legitimate liaison figure with management. More recently, Earley (1986) found that a goal-setting program initiated by a shop steward was more effective than one initiated by a supervisor in England. No such differences were found, however, in a U.S. sample. Earley (1986) concluded that English workers, as contrasted with U.S. workers, placed greater trust in their stewards than in their supervisors and, therefore, responded more favorably to a goal program sponsored by a steward than a manager. These studies suggest that the transferability of a management technique such as participative goal setting across cultural settings may be affected by prevailing work norms.

Cross-cultural differences in cultural values may, in part, explain the inconsistencies across studies concerning the effect of participation in goal setting on performance (Locke, Latham,

The authors would like to thank Rachel Grechikov for her help in the data collection for the Israeli samples, and E. A. Locke for his helpful comments on an earlier draft of this article.

Correspondence concerning this article should be addressed to P. Christopher Earley, Department of Management and Policy, University of Arizona, Tucson, Arizona 85721.

& Erez, in press). Research conducted in the United States by Latham and his colleagues (see for a brief review, Latham & Steele, 1983) has failed to demonstrate significant differences in goal acceptance and performance due to participative versus assigned goal-setting strategies. The few studies that demonstrated significant differences in performance due to participative and assigned strategies in the United States generally used tasks of sufficient complexity (e.g., Campbell & Gingrich, 1986; Earley, 1985) or difficulty that subjects did not invariably commit themselves to the goals (Earley, 1985; Erez, Earley, & Hulin, 1985). This suggests that mainly under extreme circumstances, such as when a goal is resisted or rejected (e.g., a very difficult goal), will American workers benefit from participation. Americans are characterized by a strong individualistic orientation and by a moderate to high level of power distance. These values are consistent with a system of management under which individuals will generally commit themselves to work goals assigned by their superiors (Hofstede, 1980). On the other hand, research conducted by Erez and her colleagues (Erez, 1986; Erez & Arad, 1986) in Israel, and by Matsui, Imaizumi, Onglatco, and Kaku-yama, 1985, in Japan, has shown significant effects for participation on goal acceptance and performance. A group-participation strategy is more congruent with the lower power distance, participative and collectivistic values, and with the group orientation found in Israel (Haire et al., 1966; Hofstede, 1980) than with the moderate to high power distance and individualistic orientation of American workers. The cultural differences between Japan and the United States led England (1983) to seriously question the transferability of theory Z to the United States.

Erez's (Erez & Arad, 1986; Erez et al., 1985) and Latham's (Latham & Steele, 1983) studies differed with respect to additional variables that may be related to the inconsistent effects of participation on acceptance and performance demonstrated in the literature. Goal acceptance is invariably high in Latham's but not in Erez's research. In addition, Erez uses a *tell* style of assigned goals for which individuals receive their goal without further information, whereas Latham uses a *tell and sell* style for which verbal persuasion (concerning the importance and relevance of the assigned goal) is included in the goal assignment. Finally, the goals seem to be more difficult in Erez's than in Latham's studies (see for a discussion, Locke et al., in press).

The primary purpose of the present study was to examine the relations among goal-setting strategies, goal acceptance, and performance in two countries, the United States and Israel, which differ in their cultural values. The United States is known for its individualistic values and moderate to high levels of power distance in organizations (Hofstede, 1980), and for the lack of institutionalizing employee participation programs. On the other hand, Israel is known for its collectivistic values and lower power distance, and for its employee participation programs that are institutionalized in the labor relations system (Rosenstein, 1977). The Israeli sample will be further divided into two subsamples—kibbutz (commune), and nonkibbutz (urban) students. The former are known for their extremely collectivistic values and low power distance.

On the basis of previously discussed research, the following hypotheses were proposed:

1. Participation (in goal setting) will have a positive effect on goal acceptance, particularly if goals are extremely difficult.

2. Participation will have a positive effect on performance that will be mediated by goal acceptance.

3. Cultural values will influence the effect of participation on goal acceptance and performance; the effect of participation on acceptance and performance will be greater for the Israeli than for the American subjects.

Method

Subjects

Three groups of subjects, coming from different sociocultural backgrounds, participated in the study. Two groups consisted of 120 Israeli students attending a university in Israel. One half of these students were from the kibbutz system of Israel, whereas the remaining students were from nonkibbutz, urban areas. The third group consisted of 60 American students at a midwestern university, who were randomly selected (to keep group size equal for the three groups) from a database used in the study by Erez et al. (1985). All of the students (in both countries) participated in the study on a voluntary basis, as part of a classroom exercise.

Design

The study consisted of a $3 \times 3 \times 2$ (Goal-Setting Condition \times Sample Origin \times Performance Phase) partially crossed factorial design, with performance phase as a repeated factor. The first variable, goal-setting condition, consisted of three types—assigned, representative, and participative goal setting. Goal setting by representation was implemented by having a group ($n = 5$) elect a group member to represent them in a negotiation with the experimenter concerning a goal for each group member. In the participative goal-setting condition, a group ($n = 5$) of students discussed among themselves the goal to be pursued by each group member.

The second variable, sample, consisted of three groups—Israel-kibbutz, Israel-urban, and United States. The third variable consisted of two performance phases, each lasting 20 min.

The theory tested in the present study is at an individual level, and all analyses will be conducted at the level of the individual, aggregated into the goal-setting conditions. Thus, the total number of cases was 180.

Task

Subjects in all of the conditions were asked to work on a simulated scheduling task. They were given a list of eight university courses with at least 10 different time offerings per course and were asked to assemble as many nonconflicting and nonredundant class schedules as possible using any five of the eight courses for each schedule. For the purpose of increasing experimental realism, the task directions stated that the data might be made available to the registrar's office staff for a project concerning course offerings.

All of the experimental materials (including questionnaires) were first developed in English for the U.S. subjects and then translated into Hebrew for the Israeli subjects. The translation was done independently by the experimenter and by a graduate assistant, and the two translations were found to be consistent with each other. The experimenter checked the quality of the translations during a pilot test of the experiment in Israel. Although formal procedures for back translation were not followed, the results of the pilot test confirmed the quality of the translations.

Goal-Setting Manipulations

The assigned goal setting consisted of the subjects being instructed to achieve a specific goal (10 or 25 schedules in Phases 1 and 2, respec-

tively), using a tell style of direct and brief instructions. In the representative goal-setting conditions, the groups ($n = 5$) were each asked to elect a member to represent them in a negotiation with the experimenter concerning their performance goal. The group was instructed that the goal determined by the negotiator and experimenter would be the goal for each group member. The actual negotiation was enacted in a room separate from the other group members. To begin the negotiation, the experimenter made an initial inquiry as to what the subject thought was a "difficult but attainable" goal. After hearing the subject's proposal, the experimenter gave a counterproposal. The counterproposal was based on the difference between the assigned goal and the subject's proposal. For instance, if the subject offered a goal of 8 schedules, the experimenter countered with 12 schedules (10 assigned goals + [assigned goal - subject's offer]). This procedure was used for both performance phases. Nearly all of the subjects wanted goals lower than the assigned condition target; the negotiation procedure was repeated until the target goal was reached or until it was clear that the subject would no longer negotiate.

In the participative goal-setting condition, the experimenter asked a group of subjects ($n = 5$) to decide as a group what goal would be used for each group member. It was stressed that the goal decided on by the group would be used by each group member. To begin the group's deliberations, the experimenter instructed the group members that the goal should be "difficult but attainable" and that "A pilot study has indicated that others can attain 10 (25) schedules" (the numbers given to the assigned groups for Phases 1 and 2, respectively). The experimenter used the prompt "Your goal should be difficult but obtainable . . . does this goal satisfy those constraints?" to direct the participative groups' goals to be comparable to those set for the assigned goal groups. The comments made by the experimenter were aimed at increasing subjects' willingness to set difficult goals. No other information was provided. The experimenter provided the group with a reference point, but the experimenter did not impose the goal, as can be seen from the variance in goal levels among the participative groups: Actual goals set by the representatives and participative groups ranged from 9 to 12 in Phase 1 (assigned goal was 10), and from 12 to 25 in Phase 2 (assigned goal was 25).

Experimental Measures

Goal acceptance was determined using two items: (a) "To what extent do you accept the goal?" where 1 = *strongly reject* and 7 = *strongly accept*, and (b) "How committed are you to the goal that has been set?" where 1 = *not at all committed* and 5 = *extremely committed*. The reliability (Cronbach's alpha) between these two items was .79 ($p < .01$). The responses to the two items were averaged to form a single composite goal acceptance score. Thus, the range of acceptance was from *low* (1) to *high* (6).

To better understand potential moderating effects arising in the analyses due to sample (U.S., Israel-urban, Israel-kibbutz), several dimensions of culture were assessed. The dimensions examined, collectivism and power distance, were chosen because of their theoretical connections to superior-subordinate relations such as those operating in implementing a goal-setting program (Hofstede, 1980). Collectivism refers to the extent to which an individual identifies with his or her collective, clan, or group. Power distance refers to the extent to which Superior X can influence the behavior of Subordinate Y and that Subordinate Y can influence Superior X. These two cultural values were measured using seven items based on Hofstede (1984): the first four items under collectivism and the next three items under power distance. (a) Only those who depend on themselves get ahead in life (reverse scored). (b) One should live one's life independent of others as much as possible (reverse scored). (c) Working with a group is better than working alone. (d) In society, people are born into extended families or clans who protect them in shared necessity for loyalty. (e) Powerful people should try to look less powerful than they are (reverse scored). (f) Subordinates

consider superiors as being of a different kind. (g) Other people are a potential threat to one's power and rarely can be trusted. Responses ranged on a 5-point Likert-type scale, ranging from *strongly disagree* (1) to *strongly agree* (5). The reliabilities (Cronbach's alpha) of these items were .69 and .75, for collectivism and power distance, respectively.

Procedure

The procedure used for each sample was identical; therefore, the following description refers to all of the subjects. Subjects were randomly assigned to one of three goal-setting conditions (assigned, representative, or participative). Subjects in each of the conditions ($n = 20$) were further divided into parallel groups of 5 subjects each, to keep the group size manageable. After being seated in a room with 5 subjects each, the researcher described the general purpose of the experiment, namely, an examination of how work goals influence an individual's performance of a task. It was also heavily stressed that the individuals in the study should express their honest reactions to the various stages of the experiment and that if they perceived a goal to be unreasonably difficult or highly undesirable it would be appropriate for them to reject it. Thus, the subjects were encouraged to freely reject their goals if they perceived them to be unreasonably difficult or undesirable. Next, the subjects were asked to read the task instructions. All of the subjects were given a 10-minute practice phase to familiarize themselves with the scheduling task, which they performed individually. Performance scores for the practice phase were used as a measure of a person's ability. After the practice phase, subjects underwent the goal-setting manipulations. The performance goals for the subjects were then determined according to each group's respective goal-setting condition—assigned, representative, or participative. After the goal-setting manipulation, subjects were given a brief questionnaire assessing their individual goal acceptance, and then they began to perform individually the first, 20-min phase period. After completion of the first performance phase, the goal-setting manipulations were reenacted before a second performance phase. The subjects were told that the purpose of the second phase was to assess the learning effect of goal setting on task performance under even greater time pressures. The target goal (assigned goal) for the second performance phase was 25 schedules.

On completion of the second performance phase, a questionnaire was administered to the subjects individually to determine manipulation checks, opinions about the experiment's characteristics, and personal data. Finally, the subjects were debriefed.

Results

Manipulation Checks

The amount of participation an individual perceived having experienced was measured by two items concerning the influence the subjects had during the goal setting, and his or her perceived influence relative to the experimenter. Responses to these items were made on a 5-point scale, ranging from *no influence at all* (1) to *complete control* (5). The correlation between the two items was .74 ($p < .01$). A two-way analysis of variance (ANOVA) was performed on a composite of the two items. The results of the analysis demonstrate a significant effect for goal-setting condition, but no other significant effects. A planned comparison (coding -2, 1, and 1 for the assigned, representative, and participative groups, respectively) demonstrated that the subjects in the assigned group perceived having less influence than the other goal-setting groups, $F(2, 174) = 5.38$, $p < .01$ ($M = 2.05, 2.30, 2.86$, for the assigned, representative, and participative groups, respectively).

Other checks were made for ability (using the practice phase

performance), goal level set, and task difficulty. The results of the ANOVAs demonstrated no significant effects.

Cultural Values

The mean scores on collectivism for the three samples were as follows: United States— $M = 2.32$, $SD = 0.67$; Israel—urban— $M = 2.85$, $SD = 0.81$; Israel—kibbutz— $M = 2.99$, $SD = 0.69$. The level of collectivism significantly differed across the samples, $F(2, 168) = 14.11$, $p < .05$, but not across goal-setting conditions. A post hoc contrast demonstrated that the two Israeli samples were significantly higher than the U.S. sample on collectivism, $t(54) = 4.83$, $p < .05$, but did not significantly differ from one another. No significant interaction was obtained.

The mean scores on power distance for the three samples were as follows: United States— $M = 3.35$, $SD = 0.77$; Israel—urban— $M = 2.79$, $SD = 0.87$; Israel—kibbutz— $M = 2.69$, $SD = 0.85$. The level of power distance significantly differed across the samples, $F(2, 168) = 3.22$, $p < .05$, but not across goal-setting conditions. A post hoc contrast demonstrated that the two Israeli samples were significantly higher than the U.S. sample on power distance, $t(54) = 2.08$, $p < .05$, but did not significantly differ from one another. No significant interaction was obtained.

Goal Difficulty

Approximately 62% and 9% of the subjects were able to obtain their goal in Phases 1 and 2, respectively.

Goal Acceptance

The means and standard deviations for goal acceptance, performance (adjusted using baseline ability), and goal level set for performance Phases 1 and 2 are presented in Table 1. The mean acceptance scores ranged between $M = 2.45$ and $M = 5.85$.

To test Hypothesis 1, that participation would enhance goal acceptance, a three-way repeated-measures ANOVA was conducted on goal acceptance, using phase as the repeated factor. Results are presented in Table 2. The ANOVA demonstrates a significant effect for goal-setting conditions, phase, and a Goal-Setting Condition \times Phase interaction. The interaction effect was further decomposed using a Scheffé test for post hoc contrasts (coding 1, -1 , respectively) within each goal-setting condition. The results demonstrate that goal acceptance significantly decreased for the assigned and representative goal-setting conditions from Phase 1 to Phase 2, $t(54) = 3.73, 2.78$, $p < .05$, for the assigned and representative conditions, respectively, but not for the participative goal-setting condition.

Performance

To test Hypothesis 2, that participation would increase an individual's performance, the schedules correctly assembled were examined. These data were analyzed using a repeated measures analysis of covariance (ANCOVA), using phase as the repeated factor. Prior to this analysis, the homogeneity of beta coefficients for the covariate (ability) was tested and no significant differences were found among the groups. A two-way repeated-measures ANCOVA was conducted, using ability as the

covariate and performance phase as the repeated measure. The results of the analysis (see Table 2) demonstrate a significant effect for goal-setting condition, sample, and performance phase. The Goal-Setting Condition \times Phase interaction approached significance ($p = .06$). Post hoc contrasts demonstrated that performance was significantly higher in the participative and representative than in the assigned conditions, $t(54) = 3.31$, $p < .05$ ($M = 11.21, 14.08, 14.92$, for the assigned, representative, and participative conditions, respectively, across samples and phases). Phase 2 performance was significantly higher than Phase 1 performance, and the overall performance of the kibbutz sample was the lowest of any sample. It was significantly lower than the two other samples, $t(54) = 10.97$, $p < .05$ ($M = 14.59, 14.58, 11.34$, for the U.S., Israel—urban, Israel—kibbutz samples, respectively). The near significant Goal-Setting Condition \times Phase Interaction and the pattern of means suggests that the beneficial effect of the representative and participative forms of goal setting increased across the two performance phases.

To further examine the relations among the individual variables for performance, two univariate ANCOVAs were separately conducted for Phases 1 and 2, using sample and goal-setting condition. Contrast analyses were performed between the assigned and the participative, and the representative and participative conditions (coded as 1, -1 ; and $-1, 1$, respectively). The results in both phases demonstrate that performance was significantly higher in the participative than in the assigned goal condition, $t(54) = -2.93, -3.67$, $p < .05$, for Phases 1 and 2, respectively, but the participative and representative conditions did not significantly differ for either phase. Combining the near significant interaction (Goal Setting \times Phase) with the contrast values suggests that the beneficial effect of participation increased as the goals became increasingly difficult. To more clearly interpret the various analyses, a visual display of the effect is presented in Figure 1.

Tests of Mediation

To test Hypothesis 2, that the goal-setting strategies would influence performance through the mediating effect of goal acceptance, two sets of regression analyses were conducted for each performance phase. Results are presented in Table 3.

The first set of regression analyses (3a, 3b) conducted on performance examined the effects of ability (Step 1), goal level set (Step 2), and goal-setting condition and sample (Step 3). The second set of regressions (3c, 3d) entered the variables in the same order except that goal acceptance was entered into the equation on Step 3 prior to goal-setting condition and sample (Step 4). The results for both phases demonstrate that prior to entering goal acceptance in the equation, goal-setting condition significantly predicted performance ($\Delta R^2 = .05, .08$, for Phases 1 and 2). Entering goal acceptance prior to goal-setting condition demonstrates that goal setting only explained an additional 3% of the variance ($p < .05$) in performance for either phase. Thus, the analysis provides moderate support for the hypothesized mediating effect of goal acceptance in the relation of goal-setting conditions to performance.

Tests of Moderation

Hypothesis 3 forwarded a moderating role of cultural values in the relations of goal-setting condition to goal acceptance and

Table 1
Means and Standard Deviations for Measures Across Phases 1 and 2 for Goal-Setting Condition of Sample

Measure	Goal setting								
	United States			Israel-urban			Israel-kibbutz		
	A	R	P	A	R	P	A	R	P
Phase 1									
Performance									
<i>M</i>	11.20	12.14	11.92	10.95	13.29	14.56	7.56	9.40	9.82
<i>SD</i>	2.51	4.53	6.57	5.49	3.83	3.84	3.51	2.89	3.84
Goal acceptance									
<i>M</i>	4.60	5.55	5.85	3.95	5.55	5.60	3.95	5.05	4.90
<i>SD</i>	1.76	1.88	1.75	2.03	1.47	1.27	1.85	1.79	1.86
Goal level									
<i>M</i>	10.00	10.50	13.10	10.00	11.25	11.25	10.00	10.00	10.00
<i>SD</i>	0.00	2.89	6.02	0.00	2.22	1.33	0.00	0.00	0.00
Phase 2									
Performance									
<i>M</i>	15.81	17.81	17.87	11.71	17.34	18.90	10.23	14.53	16.48
<i>SD</i>	6.69	6.78	9.52	9.83	4.31	3.46	4.20	5.09	4.45
Goal acceptance									
<i>M</i>	2.45	3.75	5.40	2.70	3.90	5.45	2.75	4.47	4.55
<i>SD</i>	1.82	2.38	1.88	1.69	2.67	1.47	1.52	2.34	2.21
Goal level									
<i>M</i>	25.00	24.25	21.90	25.00	22.50	16.50	25.00	16.00	12.75
<i>SD</i>	0.00	2.85	3.65	0.00	3.36	2.35	0.00	2.41	1.52

Note. A = assigned goal condition; R = representative goal condition; P = participative goal condition. *n* = 20/cell; means adjusted using ability as a covariate.

performance. Hypothesis 3 was explored by tests of moderation, using hierarchical regression analysis for Phase 2 of the experiment. The second phase was chosen because the relative spread between conditions was the largest during this phase.

Table 2
Analyses of Variance (Covariance) for Goal Acceptance and Performance

Source	<i>df</i>	<i>MS</i>	<i>F</i>	η^2
Goal acceptance				
Goal setting (GS)	2	112.93	21.90*	.20
Sample (SPL)	2	3.15	0.61	.00
Phase (PHS)	1	102.48	55.33*	.25
GS \times SPL	4	3.58	0.69	.00
GS \times PHS	2	12.98	7.01*	.06
SPL \times PHS	2	4.13	2.23	.02
GS \times SPL \times PHS	4	1.46	0.79	.00
Performance ^a				
Goal-setting (GS)	2	348.10	9.65*	.08
Sample (SPL)	2	373.40	10.35*	.09
Phase (PHS)	1	1725.17	100.69*	.37
GS \times SPL	4	52.12	1.45	.00
GS \times PHS	2	48.94	2.86	.02
SPL \times PHS	2	45.04	2.63	.02
GS \times SPL \times PHS	4	9.57	0.56	.00

^a Using baseline performance (ability) as a covariate.
* *p* < .01.

The procedure for a moderation test using hierarchical regression analysis is described in detail elsewhere (James & Brett, 1984). In the present study, three regression analyses each were conducted on Phase 2 performance and goal acceptance, entering in the following order: ability (performance analysis only), goal-setting condition (coded -2, 1, 1, for the assigned, representative, and participative conditions, respectively) and the hypothesized moderator (sample, collectivism, or power distance), and Goal-Setting Condition \times Moderator interaction term. A moderator effect is demonstrated by a significant interaction term. The results of the regression analyses demonstrate a significant moderator effect on performance for sample and power distance ($\Delta R^2 = .05, .04, p < .05$, and *t*[for beta] = -2.71, -1.92, *p* < .05, for Goal Setting \times Sample and Goal Setting \times Power Distance, respectively) but not for collectivism ($\Delta R^2 = .00, ns$). The direction of the interaction term and an examination of the means indicates that participative and representative goal setting had stronger effects on performance for individuals low in power distance (as in Israel) than high in power distance (as in the U.S.). No significant moderator effects on goal acceptance were obtained, although the interaction term for Goal Setting \times Power Distance approached significance, *t*(for beta) = -1.77, *p* = .08.

To better understand how the moderators relate to goal acceptance and performance, specific contrasts (comparing the assigned with the representative and participative goal-setting conditions, coded as -2, 1, 1, respectively) within each country for Phase 2 of the experiment were examined. The results demonstrate that performance and goal acceptance were signifi-

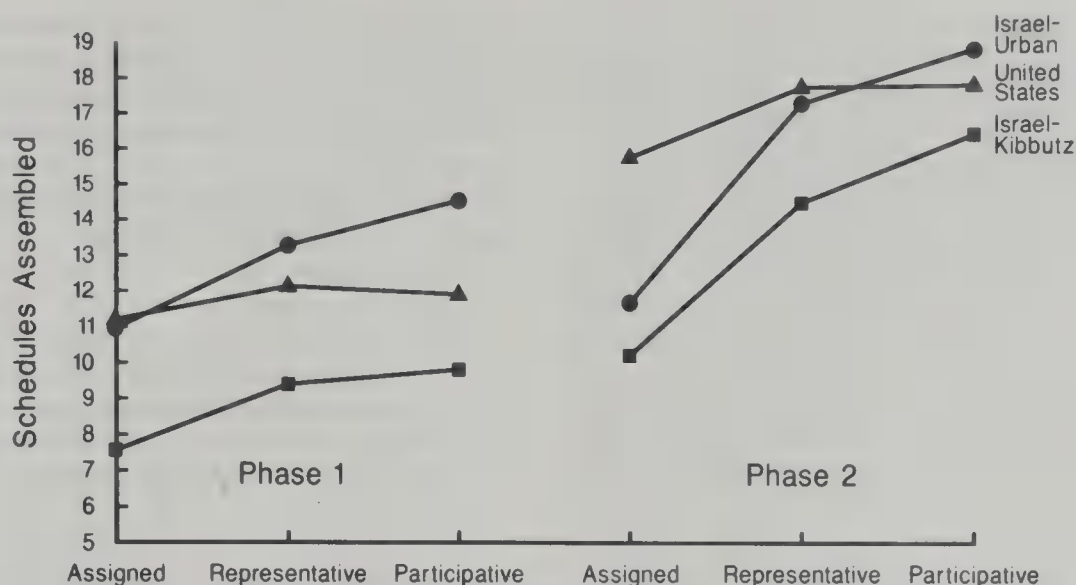


Figure 1. Presentation of adjusted performance means across goal-setting strategies and countries for Phases 1 and 2.

cantly higher in the participative and representative than in the assigned goal-setting condition in both Israeli samples, $t(54) = 2.99, 3.21$, for performance, and $2.70, 3.39$, $p < .05$, for goal acceptance, for the Israeli-kibbutz and Israeli-urban samples, respectively. The contrast analysis for performance was not significant for the U.S. sample, but it approached significance, $t(54) = 1.66$, $p = .09$; the contrast for goal acceptance was significant for the U.S. sample, $t(54) = 4.57$, $p < .05$.

Further exploratory contrasts were conducted on performance comparing individuals' (Phase 2) performances within each of the types of goal setting (coded $-1, 1$, for the U.S. vs.

Israeli-kibbutz, and, $-1, 1$ for the U.S. vs. Israeli-urban, respectively). The results demonstrate no significant differences among samples for the representative and participative goal-setting conditions, but there was a significant difference for the assigned condition, $t(50) = -2.47$, $p < .05$, -1.52 (*ns*), for the United States versus Israeli-kibbutz, United States versus Israeli-urban, respectively. The U.S. sample performed significantly better than the Israeli-kibbutz but not the Israeli-urban samples in the assigned goal condition.

Taken together, these analyses suggest that goal acceptance significantly increased with participation in all samples. Hence, cultural values only seem to weakly moderate the effect of goal-setting strategies on goal acceptance (reflected by the marginal effect for the Goal Setting \times Power Distance interaction). Performance, however, is differentially affected by goal setting across countries and levels of power distance; assigned goals are relatively more effective in a high than in a low power distance culture. The participative and representative forms of goal setting appear to be more effective than assigned goal setting for the Israeli samples (low power distance) but not for the U.S. sample (moderate to high power distance). Although the mean performance levels for the U.S. sample in both phases were higher in the representative and participative than in the assigned goal-setting conditions, these differences were not statistically significant (the contrasts were significant at $p = .15, .09$, for Phases 1 and 2, respectively).

Discussion

The main purpose of the present study was to examine the effect of cultural values on the relations among goal-setting strategies, goal acceptance, and performance for moderately difficult and extremely difficult goals. Specifically, the effect of the Israeli and the U.S. cultures, which differ in their cultural values, was examined.

Consistent with previous research, goal acceptance was significantly affected by goal difficulty (Erez & Zidon, 1984) and by participative versus nonparticipative strategies (Earley,

Table 3
Hierarchical Regression Analyses for Performance

Variable entered	Step	R^2	R^2 change (per step)	T (per variable)
Phase 1 (A)				
Ability	1	.17	.17	5.84**
Goal level	2	.18	.01	-0.76
Goal setting	3	.23	.05	3.22**
Sample	3			0.68
Phase 2 (B)				
Ability	1	.04	.04	2.66**
Goal level	2	.06	.02	-1.82
Goal setting	3	.14	.08	3.42**
Sample	4			-0.89
Phase 1 (C)				
Ability	1	.17	.17	5.84**
Goal level	2	.18	.01	-0.76
Goal acceptance	3	.26	.08	3.96**
Goal setting	4	.29	.03	2.11*
Sample	4			1.13
Phase 2 (D)				
Ability	1	.04	.04	2.66**
Goal level	2	.06	.02	-1.82
Goal acceptance	3	.17	.11	4.58**
Goal setting	4	.20	.03	2.15*
Sample	4			-0.90

* $p < .05$. ** $p < .01$.

1985; Erez & Arad, 1986; Erez et al., 1985; Hannan, 1975). The highest acceptance level was obtained in the participative condition and the lowest in the assigned goal condition. Moreover, acceptance in the assigned and representative conditions was more strongly affected by the extremely difficult goals and it significantly dropped between Phases 1 and 2. Yet, acceptance in the participative condition remained stable and did not even drop significantly for extremely difficult goals. This result suggests that participation helps to maintain high acceptance even for the most difficult goals. Goal acceptance intervened in the relation between participation and performance. When acceptance was statistically controlled, participative goal setting accounted for very little variance in performance. Hence, the two-step model postulated by Erez et al. (1985) has received further support.

The effect of participation on acceptance was significant in all three samples. Hence, culture does appear to, at best, act as a weak moderator for the effect of participation. The inconsistency between Latham's (Latham & Steele, 1983) studies, which showed no significant effect of participation on goal acceptance, and Erez's (Erez & Arad, 1986; Erez et al., 1985) studies, which found a significant effect, cannot be explained by cultural differences alone. Other factors should, therefore, be examined to explain the differences. For example, the high variance in acceptance found in Erez's studies versus the lack of variance reported in Latham's studies may explain the differences in the effects. Other factors could be as follows: the tell style used by Erez for assigning the goals versus the tell and sell style used by Latham, the unimportant versus important tasks used (task meaningfulness), and the extremely difficult versus nondifficult goals in Erez's and Latham's studies, respectively. These factors are currently being examined in a series of experiments jointly conducted by Erez, Latham, and Locke, and will be reported elsewhere.

The effect of goal-setting strategies on performance was significant in both phases of the experiment. Performance was lower in the assigned than in the representative and participative conditions. The goal-setting effect interacted somewhat ($p = .06$) with goal difficulty, and indicated that the differences between the assigned and participative conditions became more evident as goal difficulty increased in Phase 2.

The moderating effect of culture on the relations between goal-setting treatments, acceptance, and performance should be discussed in light of three research evidences. First, the omnibus interaction effect of sample with goal-setting conditions across both performance phases was not significant, and did not provide direct support for the moderating effect. However, the post hoc comparisons in Phase 2 only demonstrated significant moderating effects for both sample and power distance. The U.S. sample outperformed the Israeli samples within the assigned goal-setting treatment. No such differences were found in the representative and participative goal-setting treatments. This finding proposes that the more collectivistic and lower power distance Israeli subjects adversely reacted to the nonparticipative, assigned goals as compared with the more individualistic and higher power distance American subjects.

Second, the urban-Israeli subjects in the participative condition performed as well as the U.S. subjects, even though the latter had higher goals. The mean level of goal difficulty in Phase 2 for the U.S. samples was 24.2 and 21.9, for the participative

and representative goal conditions, respectively, whereas for the Israeli samples it was 18.2 and 14.6. This finding suggests that, relative to their goals, the Israeli samples performed better than the U.S. sample in the participative goal-setting conditions in Phase 2.

Third, in the assigned goal condition, goal acceptance was more strongly related to performance for the two Israeli than for the U.S. sample ($r = .10, .41, .58$, for the U.S., Israeli-urban, Israeli-kibbutz samples, respectively). Thus, some aspect of culture moderated the relation between acceptance and performance. The high relation between the motivational response of acceptance and the behavioral response of performance for the Israeli rather than the American subjects seems to reflect some type of cultural differences. One possible explanation lies in the effects obtained for power distance. Americans have a higher power distance than the Israelis (Hofstede, 1980), and therefore, they are more likely to acquiesce and use the goals they are assigned by their superiors, even though they may not accept (or be committed to) them. Another possible explanation is that the Americans are more "masculine" than the Israelis (Hofstede, 1980), and are less likely to express their emotions in their behavior. (Masculinity refers to the extent to which individuals in a society openly express their emotions, compete with one another, and open themselves to others.) The preceding discussion supports the argument that culture, in some fashion, moderates the relation between goal-setting conditions and performance.

The results lead to the conclusion that the differences between the cultures under study are not so much in terms of the effects of participation as they are in terms of the reaction of individuals to their assigned goals. In line with this argument, we suggest that the disagreement between Latham's and Erez's studies is on the effect of the assigned goals rather than on the effect of participation on performance. Vroom and Yetton (1973) proposed that participation should be preferred to nonparticipation when the level of acceptance is low. In Latham's studies, acceptance was similarly high in both assigned and participative goal-setting treatments; thus, no differences in performance were found. In the present study, acceptance was significantly lower in the assigned than in the participative goal-setting conditions for both the Israeli and the American sample. However, only in the Israeli sample was acceptance highly related to the performance of assigned goals.

Culture as a moderator supports the contingency approach, which proposes that cross-cultural research can be advanced by relating the level of cultural values to management practices (Wilpert, 1974). A contingency approach is useful for predicting the effect of different managerial practices across cultures. The match between culture and managerial style becomes crucial in the context of multinational corporations. Decisions on what types of managers to appoint to the organization outside the home country and whether to develop a local cadre of managers or to send expatriates is affected by the congruence between the values of the host country and managerial style espoused by the home corporation. Miles and Snow (1984) portrayed the future organization as one with no geographical borders. Different stages of the production process will be performed in different countries. Hence, a contingency model will be crucial for the successful implementation of management practices. Today, there is a growing desire of managers to adopt

Japanese management practices and to develop cultural rather than bureaucratic controls (Jaeger, 1983). However, there are some unique characteristics of the Japanese culture and norms in the workplace that do not exist in the U.S. culture, for example, and whether they will be effective in the United States without major modification is questionable (England, 1983). The contingency approach should be considered when transferring management techniques across cultures.

References

- Campbell, D. J., & Gingrich, K. E. (1986). The interactive effects of task complexity and participation on task performance: A field experiment. *Organizational Behavior and Human Decision Processes*, 38, 162-180.
- Coch, L., & French, J. R. P. (1948). Overcoming resistance to change. *Human Relations*, 1, 512-532.
- Earley, P. C. (1985). The influence of information, choice, and task complexity upon goal acceptance, performance, and personal goals. *Journal of Applied Psychology*, 70, 481-491.
- Earley, P. C. (1986). Supervisors and shop stewards as sources of contextual information in goal setting: A comparison of the U.S. with England. *Journal of Applied Psychology*, 71, 111-118.
- England, G. W. (1983). Japanese and American management: Theory Z and beyond. *Journal of International Business Studies*, 14, 131-141.
- Erez, M. (1986). The congruence of goal-setting strategies with socio-cultural values, and its effect on performance. *Journal of Management*, 12, 585-592.
- Erez, M., & Arad, R. (1986). Participative goal setting: Social, motivational, and cognitive factors. *Journal of Applied Psychology*, 71, 591-597.
- Erez, M., Earley, P. C., & Hulin, C. (1985). The impact of participation upon goal acceptance: A two-step model. *Academy of Management Journal*, 28, 50-66.
- Erez, M., & Zidon, I. (1984). Effect of goal acceptance on the relationship of goal difficulty to performance. *Journal of Applied Psychology*, 69, 69-78.
- French, J. R. P., Israel, J., & Ås, D. (1960). An experiment in a Norwegian factory: Interpersonal dimension in decision-making. *Human Relations*, 13, 3-19.
- French, J. R. P., Kay, E., & Meyer, H. H. (1966). Participation and appraisal system. *Human Relations*, 19, 3-19.
- Haire, M., Ghiselli, E. E., & Porter, L. W. (1966). *Managerial thinking: An international study*. New York: Wiley.
- Hannan, R. L. (1975). *The effects of participation in goal setting on goal acceptance and performance*. Unpublished doctoral dissertation, University of Maryland.
- Heller, F. A., & Wilpert, B. (1981). *Competence and power in managerial decision making*. Chichester, NY: Wiley.
- Hofstede, G. (1980). *Culture's consequences: International differences in work related values*. Beverly Hills, CA: Sage.
- Hofstede, G. (1984). *Culture's consequences: International differences in work related values* (abridged). Beverly Hills, CA: Sage.
- IDE (Industrial democracy in Europe—International research group). (1981). London: Oxford University Press.
- Jaeger, A. (1983). The transfer of organizational culture overseas: An approach to control in the multinational corporation. *Journal of International Business Studies*, 14, 91-114.
- James, L. R., & Brett, J. M. (1984). Mediators, moderators, and tests for mediation. *Journal of Applied Psychology*, 69, 307-321.
- Latham, G. P., & Steele, T. P. (1983). The motivational effects of participation versus goal setting on performance. *Academy of Management Journal*, 26, 406-417.
- Locke, E. A., Latham, G. P., & Erez, M. (in press). The determinants of goal commitment. *Academy of Management Review*.
- Locke, E. A., & Schweiger, D. M. (1979). Participation in decision-making: One more look. In B. M. Staw (Ed.), *Research in organizational behavior* (Vol. 1, pp. 265-340). Greenwich, CT: JAI Press.
- Matsui, T., Imaizumi, T., Onglatco, M. L., & Kakuyama, T. (1985). *Effects of individual versus group goals and feedback on task performance*. Unpublished manuscript.
- Maurice, M., Sorge, A., & Warner, M. (1980). Societal differences in organizing manufacturing units: A comparison of France, West Germany, and Great Britain. *Organization Studies*, 1, 59-86.
- Miles, R. E., & Snow, C. C. (1984). Fit, failure, and the Hall of Fame. *California Management Review*, 26, 10-28.
- Ronen, S., & Shenkar, O. (1985). Clustering countries on attitudinal dimensions: A review and synthesis. *Academy of Management Review*, 3, 435-454.
- Rosenstein, E. (1977). Worker participation in management: Problematic issues in Israeli system. *Industrial Relations Journal*, 8, 55-69.
- Tannenbaum, A. S., Kavcic, B., Rosner, M., Vianello, M., & Wieser, G. (1974). *Hierarchy in organizations*. San Francisco: Jossey-Bass.
- Vroom, V. H., & Yetton, P. W. (1973). *Leadership and decision-making*. Pittsburgh, PA: University of Pittsburgh Press.
- Wilpert, B. (1984). Participation in organizations: Evidence from international comparative research. *International Social Science Journal*, 36, 355-366.

Received July 24, 1986

Revision received March 17, 1987

Accepted March 30, 1987 ■

Effect of Values on Perception and Decision Making: A Study of Alternative Work Values Measures

Elizabeth C. Ravlin and Bruce M. Meglino

Riegel and Emory Human Resource Research Center, College of Business Administration
University of South Carolina

Four alternative methods of measuring values were used to examine the impact of work values on perception and decision-making tasks. Perception and its relation to values was assessed using interpretation of ambiguous stimuli. The effect of values on decision making was evaluated using within-subject regression analyses of 20 separate decisions. A total of 103 undergraduate subjects completed values measures and the perceptual and decision-making tasks in three work sessions, each separated by from 2 to 4 days. A rank order measure of values related more consistently to perception and decision making than did other measurement methods. Results also provide some support for a theory of values in which values affect perceptual organization and act as a guide to decision making.

Values have traditionally played an important role in the understanding of job satisfaction, emotion (see Lazarus & Folkman, 1984; Locke, 1976), and the behavior of individuals at work (England, 1967). Recently, however, investigators have become concerned that the values responsible for the success of the American system (the work ethic in particular) may be changing or eroding (Cherrington, 1980; Spence, 1985; Yankelovich, 1979). Individual values have also been identified as manifestations of organizational culture (Schein, 1985), which in turn, has been linked to an organization's success (Deal & Kennedy, 1982; Peters & Waterman, 1982). Finally, the accomplishments of other systems, particularly the Japanese, are believed by some researchers and practitioners to be caused by Japanese managers' and employees' possession of a more appropriate set of values than those held by individuals in other systems (Howard, Shudo, & Umeshima, 1983; Ouchi, 1981).

This cross-cultural research has severely challenged the assumption of hedonism, which is the basis of most models of work motivation. Specifically, Staw (1984) observed that although "Western models of motivation emphasize individual gain and self-interest, the Japanese system relies more heavily on motivation for collective welfare and appears to be more altruistically based" (p. 651). Because individual values are important constructs for understanding behavior not directly based on striving to maximize pleasure or individual gain (Locke, 1975), they may well assume a more prominent role in future theories of motivation (see, e.g., Locke & Henne, 1986).

Because values have traditionally been employed in a variety of ways, any study of values must be concerned with a number of interrelated issues. Specifically, one must consider the nature of values themselves, how values affect individuals, how values

are to be measured, and the specific values to be studied. One obvious concern is the relation of the nature of values to the model underlying their measurement. Values as motivational elements are generally conceptualized in two ways, reflected by different measurement methods. The first view defines values as hierarchical in nature (Locke, 1976, 1982; Rokeach, 1973) and, thus, requires a within-subject ipsative design for measurement. This approach implies that individuals have a preference ordering or ranking of values to which they refer in making behavioral choices. Allport, Vernon, and Lindzey (1960), and Rokeach (1973), used this ipsative approach in which each value is measured at the expense of the others. Such ipsative measures, however, are subject to substantial analytic limitations that reduce their flexibility in making between-subject comparisons (Hicks, 1970). Because the score for each value is determined by the scores for all other values, correlations between value scores are meaningless (in the case of two subscales, $r = -1$). Correlations with other variables are also constrained by the deterministic relation between value subscales. Hicks (1970) has made the point that transformation of such ipsative scores to normative measures is impossible, although normative scores can be used ipsatively.

The second general view does not *necessarily* conceptualize values as hierarchically organized. This view allows for variance in the importance of individual values, and also in the importance of the full array of values held, or the total importance of values to a person. If this total importance of values is significantly different between individuals, the values measurement method must be flexible enough to take this into account. Measures reflecting this conceptualization, including England's (1975) Personal Values Questionnaire (PVQ), have the additional advantage of facilitating intersubject comparisons, which ipsative measures, when properly used, do not. Despite the fact that this approach to understanding work values allows for a more complex conceptualization of values as possibly equal in motivating power or cued by situational considerations, a hierarchical interpretation, which in fact dictates ipsative measures

Preparation of this article was supported in part by the Riegel and Emory Human Resource Research Center.

Correspondence concerning this article should be addressed to Elizabeth C. Ravlin, College of Business Administration, University of South Carolina, Columbia, South Carolina 29208.

as the appropriate ones, is used almost exclusively in conjunction with this type of measure.

In addition to their cognitive organization, another conceptual aspect of values is inextricably linked to their measurement. Although there are different types of values, social values, which are studied here, represent *general* modes of behavior that an individual "should" or "ought" to exhibit (Fallding, 1965; Kluckhohn, 1951; Rokeach, 1973; Schein, 1985; Williams, 1968). This "oughtness" characteristic—perhaps the most distinctive feature of these values—arises from the social nature of the construct itself. Although all values may be initially established on the basis of experiences of pleasure and pain (Locke, 1975), social values come to represent the efforts of a social system (e.g., group, organization, society) to impose concern for the welfare of the system as a whole on the actions of its individual members (see Kluckhohn, 1951, for a discussion of this point). Thus, a society socializes individuals to value "honesty" because it is important to the society as a whole that individuals do not cheat or steal. Social values are, therefore, phenomena that are highly socially desirable and, as such, tend to be strongly endorsed by all individuals. This is another powerful argument for the use of an ipsative measurement model. If a number of values are measured independently of each other, scores for each of the values may be artificially inflated, reducing any differences between scores and making it difficult to predict behavior that involves a choice among values (e.g., whether to strive to achieve success at the expense of helping others). It may also be extremely difficult to separate what is essentially an individual's endorsement of a socially desirable phenomenon from what is an error caused by the respondent giving a socially desirable response. Forced choice scales have been used to control in part for social desirability (Anastasi, 1982); however, their use may create another problem. This type of measure may fail to capture the essence of values themselves (i.e., social desirability).

Another issue of critical importance concerns the impact of values on individuals. Two distinct processes are thought to be involved in linking values to choice behavior (England, 1967, 1975). One process, which has rarely been examined (for an exception see Postman, Bruner, & McGinnies, 1948) concerns perceptual organization. Values are thought to influence the selection and interpretation of external stimuli, thereby affecting the organization of behavioral choices or the formulation of alternative courses of action. In addition to this indirect influence, values are thought to guide action in a direct manner by influencing behavioral choice (e.g., Rokeach, 1973). Values, therefore, should relate to the organization or interpretation of stimuli as well as to actual decision-making behavior. Our study examined these theoretical relationships by assessing the degree to which subjects' values, as measured by various measurement methods, related to their perceptions on a pseudorecognition task and decisions on a judgment task. As with other central constructs (see Epstein, 1979, 1980), values should predict broad modes of behavior over time. Consequently, at any point in time, the relation between values and perceptions on the pseudorecognition task, and values and decisions on the judgment task, was expected to be somewhat weak.

The influence of values on perception and behavioral choice may, however, be enhanced by situational moderators. For ex-

ample, prevalent social norms and situational uncertainty are thought to be key elements influencing the perceptual and choice processes. We investigated the impact of uncertainty in the form of ambiguous stimuli and will discuss it ahead.

The final issue of importance to a study of values concerns the number and type of values investigated. A large number of values studies have considered only one value, which makes it impossible to address the question of how a network of values affects perception and decision making. It is also impossible to examine whether measurement models should be hierarchical. Thus, we examined a number of values that were chosen on the basis of an earlier phase of this research. In that phase, Cornelius, Ullman, Meglino, Czajka, and McNeely (1985) used a variant of Flanagan's (1954) critical incident technique in a survey of 966 employees at different levels in a variety of organizations throughout the United States. These employees completed a one-page questionnaire asking them to focus on an individual they knew well at work and to identify one value that that person held about life in general. They were then asked to describe an incident at work that illustrated why they felt that the person held that particular value. These behavioral incidents were then sorted into separate value categories by six independent expert judges. The categories that accounted for a substantial number of incidents (greater than 5% each) were as follows: achievement (19.9%), concern for others (15.1%), honesty (11.6%), working hard (10.7%), positive outlook (10.7%), helping others (10.0%), and fairness (5.8%). For the present study, the achievement and working hard categories were combined (called *achievement*), as were the concern for others and helping others categories (called *helping*). These two categories, plus *honesty* and *fairness*, formed the basis for the values measured. The positive-outlook category was not included because it was judged to be the least relevant to prior work values research.

Measures

Four alternative methods were used to measure the values identified above. All of them used direct questions, as opposed to alternative methods such as projective techniques (see Miner, 1977), or observation of action patterns. The measures consisted of a simple rank ordering of the four values (rank measure), assignment of a fixed number of points among the four values (point-assignment measure), a forced-choice measure based on a series of behavioral incidents related to each of the four values (forced-choice measure), and ratings on item response scales of the same behavioral incidents (Likert-summed scales measure). The Likert measure, because it allowed values to be measured independently of each other, was nonipsative and therefore contained no control for social desirability. The rank, point-assignment, and forced-choice measures were purely ipsative and provided some control for socially desirable responses. Note, however, that these techniques do not necessarily eliminate such bias. For example, Anastasi (1982) observed that forced-choice measures, in particular, may only partially eliminate the effects of social desirability.

To evaluate each measurement method in terms of its susceptibility to the effects of social desirability response bias, we collected scores on the Marlowe-Crowne Social Desirability Scale

(Crowne & Marlowe, 1964). We also used two commonly used measures of single values to gain further information concerning the construct validity of the measures. The Pro-Protestant Ethic subscale (Blood, 1969) indicates the degree to which a person believes that individual worth is the result of work or achievement in the work domain, whereas Crandall's (1975) Social Interest Scale (SIS; Crandall, 1977; Crandall & Harris, 1976) measures interest in and concern for others.

Hypotheses

Predictions were made regarding the relation of the various values measures to perception, decision making, response bias, and other measures. They are stated as the following hypotheses:

1. Subjects' responses on the ranking, point-assignment, and forced-choice measures will be related to the frequency with which they interpret ambiguous stimuli as representing the four value categories on a perceptual task. Because value schemas actively influence perception and interpretation, stimuli will be more likely to be interpreted as indicating more dominant values. The influence of social desirability response bias should reduce the relation between Likert scale responses and frequency responses.

2. Subjects' responses on the ranking, point-assignment, and forced-choice values measures will relate positively to the weights they place on the values in making a decision. Behavioral channelling by values should influence subjects to make decisions in a manner congruent with their values. The relation between decision weights and Likert scale responses will be affected by social desirability response bias and therefore is not predicted to be significant.

3. Responses on the Likert measure will be more highly related to responses on the Marlowe-Crowne Social Desirability Scale than will responses on rank, point-assignment, and forced-choice measures. Although ranking, point-assignment, or forced-choice methods provide control for an individual's desire to appear socially attractive, Likert items allow responses to be inflated.

4. Subjects' orderings of achievement should relate positively to their score on the Blood Pro-Protestant Ethic subscale. Subjects who rate achievement high, as compared with helping, honesty, and fairness, should score higher on the Pro-Protestant Ethic subscale. This relation should be stronger for the ranking, point-assignment, and forced-choice measures than for the Likert measure because social desirability response bias will affect Likert scale responses.

5. Subjects' orderings of helping should relate positively to their score on Crandall's SIS. Subjects who rate helping high, as compared with achievement, honesty, and fairness, should score higher on the SIS. This relation should be stronger for the ranking, point-assignment, and forced-choice measures than for the Likert measure because social desirability response bias will affect Likert scale responses.

Method

Development of Values Instruments

Development of the forced-choice, Likert, rank, and point-assignment methods is described in this section.

Forced-choice measure. For each of the four value categories, representative behavioral incidents were selected from those gathered by Cornelius et al. (1985). These incidents were edited into single-sentence items describing a specific behavior (e.g., achievement: "doing whatever work is required to advance in your career"). Additional items were also written to generate 25 items for each value. These items were developed by substitution of synonyms, changes in referent and intensity, and shifts in gender emphasis in the original behavioral incidents. The 100 items were then rated for desirability and for the extent to which each represented the value category in question. This was accomplished using two separate surveys, each administered to a different group of undergraduate students.

In the first survey, we asked 100 students to rate the desirability of each item on a 5-point scale, ranging from *not desirable* (1) to *highly desirable* (5). This was done to obtain statements matched in desirability for the forced-choice measure. Choosing between two equally desirable behaviors should control for most bias toward socially acceptable responding. In the second survey, we asked a separate group of 99 students to indicate on a 5-point scale, ranging from *not at all* (1) to *a very great extent* (5), the extent to which each item related to a specific definition of its value category. This survey was taken to ensure that only items that were representative of their value category to the subject population were included in the forced-choice measure.

On the basis of these surveys, 48 items were selected for inclusion in the forced-choice values questionnaire. Of these items, 31.25% were from the original behavioral incidents and 68.75% were generated for this study, modeled on the original incidents. To be selected, an item had to receive an average rating of 3.5, or above, on the extent to which it reflected its value category (3 = *a moderate extent*). In addition, no items were included that received significantly different responses on social desirability from men and women. This second criterion was imposed to eliminate sex bias in the resulting survey. The final, and most stringent, criterion for selection was that each item had to pair with another item representing a different value, such that there existed no significant difference between the items in social desirability.

The resulting questionnaire contained 24 equally desirable pairs of items. Desirability ratings appear in Table 1. Each item in a pair reflected one of the four different values. Pairing was done systematically such that a different item reflecting a particular value appeared in 12 pairings. The questionnaire asked subjects to indicate, on a 9-point scale, how much emphasis they should place on one item in each pair as opposed to the other item in the pair. Because each item was scored at the expense of the other in the pair, this forced-choice questionnaire yielded purely ipsative scores.

Likert measure. A second questionnaire was also constructed on the basis of the same items contained in the forced-choice questionnaire described earlier. In this questionnaire, each of the 48 individual items was evaluated on a 5-point scale, ranging from *never* (1) to *always* (5) that reflected how often the respondent should engage in each of the behaviors described. Because each item was rated independently of the others, scores on this questionnaire were not ipsative in nature.

Rank and point-assignment measures. We developed two additional values measures for this study. In the first, we provided subjects with a definition of each value (the same definitions originally used to evaluate the extent to which each item related to its specific value) and asked them to rank the values according to how they felt each should be emphasized in their behavior (1 = greatest emphasis). This format is very similar to one used by Rokeach (1973). In the second measure, we provided subjects with the same value definitions but asked them to allocate a total of 25 "points" across all four values according to how they felt each should be emphasized in their behavior (more points = greater emphasis).

The procedures we have described yielded four different measures of the same four values: a forced-choice questionnaire that asked respon-

Table 1
Mean Desirability Ratings for Pairs of Behavioral Incidents
Representing Different Values

Pair	Value			
	Achievement	Helping	Honesty	Fairness
1	3.90	—	—	3.97
2	—	3.98	3.98	—
3	—	4.04	—	4.09
4	4.12	—	4.11	—
5	—	—	3.83	3.89
6	3.90	3.88	—	—
7	—	4.22	4.18	—
8	—	4.13	—	4.14
9	4.16	—	4.15	—
10	4.17	4.17	—	—
11	—	—	4.18	4.17
12	4.16	—	—	4.15
13	—	4.28	—	4.27
14	4.26	—	4.21	—
15	4.29	4.28	—	—
16	—	—	4.32	4.31
17	4.33	—	—	4.33
18	—	4.32	4.33	—
19	4.18	—	4.23	—
20	4.41	4.42	—	—
21	—	—	4.39	4.43
22	4.18	—	—	4.11
23	—	4.10	4.13	—
24	—	4.26	—	4.25

Note. Data reported in this table are based on a sample of 100 undergraduate subjects responding to a 5-point scale. *M* overall desirability rating for achievement = 4.172, helping = 4.173, honesty = 4.170, and fairness = 4.176.

dents to indicate their relative emphasis between 24 pairs of behavioral statements, a Likert questionnaire that allowed independent ratings of 48 behavioral statements according to how often the behavior should be performed, a rank ordering of the four values, and a point allocation among the four values. The forced-choice, rank, and point-assignment measures are purely ipsative, whereas the Likert measure is nonipsative. Because it is inappropriate to use ipsative scores for intersubject comparisons, the Likert scores were treated as ipsative scores for purposes of comparison with the other methods of measuring values.

The instructions for each of the measures asked subjects to respond in terms of the way they felt they *should* or *ought* to behave, rather than asking for actual behavior or for an endorsement of phrases or adjectives as most other values measures do. This method is consistent with the commonly accepted definition of social values. It is also consistent with the expectation that values, as central constructs, should be relatively stable, should predict broad modes of behavior over time, and should have a relatively weak relation to a specific behavior at a given point in time.

Subjects and Procedures

Subjects for this study were 103 undergraduate students recruited from junior- and senior-level management courses, who participated for extra credit. Subjects participated in three 45-min data-gathering sessions, each separated by from 2 to 4 days. Questionnaire data were gathered during the first two sessions, which were counterbalanced to control for order effects. At one session, the forced-choice, Protestant Ethic Scale, point-assignment, and SIS measures were administered. At the

other session, the Likert, Marlowe–Crowne, and rank measures were collected. The perceptual and decision criterion tasks were both completed during the third session and are described next.

Perceptual task. The first criterion task was a pseudorecognition exercise that has been used in a number of other studies (Cottrell, Wack, Sekerak, & Rittle, 1968; Henchy & Glass, 1968; Zajonc & Nieuwenhuyse, 1964; Zajonc & Sales, 1966) to assess an individual's hierarchical ordering of responses. A series of 25 nonsense words were flashed on a screen using a Gerbrands Model 300-C digital millisecond timer and a Gerbrands Model 66 shutter. The shutter speed was set, on the basis of pretests, such that subjects were unable to recognize the actual word. They were told that although they would not actually see the word, their minds would absorb its subliminal image. They were then asked to guess at the value category that was most closely related to each word. In prior studies, responses corresponded to the relative dominance of the concepts in a subject's cognitive hierarchy. The actual number of guesses a subject made in a value category was recorded as the frequency score for that value.

Decision task. The second criterion task required subjects to simulate a manager in making 20 separate overall evaluation decisions, ranging from *outstanding* (1) to *poor* (7), for 20 fictitious employees on the basis of each employee's work record. Each work record was presented in the form of a profile that contained an indication of the employee's performance on four dimensions corresponding to each of the four values of interest, using the same 7-point scale, ranging from *outstanding* (1) to *poor* (7), that subjects were to use in their overall evaluation. Subjects were told that the ratings had been provided by a supervisor and that these ratings should form the basis of the overall evaluation. Thus, the task provided a simulation of a common managerial activity, evaluating employees, using ratings provided by others. The value ratings were systematically varied across the fictitious profiles so that subjects evaluated profiles representing all possible combinations of high (2 = substantially above average) and low (5 = slightly below average) performance on the value dimensions (16 profiles) plus 4 duplicate profiles. For example, a subject would evaluate an individual who had obtained a rating of 2 on fairness, 2 on honesty, 5 on helping, and 2 on achievement. This procedure provided complete independence between independent variable ratings for the 16 unique profiles, thus aiding value-weight interpretation. (Including the duplicate profiles resulted in interrating correlations ranging between $-.1$ and $.1$, with a median value of 0 .) Within-subject regression analysis was then used to estimate the emphasis each subject placed on each value in making his or her evaluation. The independent variables entering these equations were the four value ratings assigned by the researchers to each of the 20 fictitious employees. The dependent variable was the subject's overall evaluation. The four standardized beta weights obtained for each subject indicated the emphasis placed on each of the four values by the subject in making the overall evaluation.

Results

Because the ipsative nature of most of the data precludes use of the usual between-subject statistics, we compared rank orderings of the values as produced by the measuring instruments and criterion procedures (the perceptual and decision tasks). To obtain appropriate within-subject rankings from the Likert measure, we adjusted subject's scores to compensate for the relative differences in mean responses across the four values for this population. Specifically, the score for each subject on each value was adjusted by the mean of the population on that particular value to provide a standard scale of response. This procedure corrected the ordering of within-subject ranks on the basis of responses of the population under study by equating across-subject means for all four values.

Table 2
Relation Between Values Measures and Task Criteria

Measure	Perceptual task			Decision task		
	Z ^a	Proportion significant ^b	r ^c	Z ^a	Proportion significant ^b	r ^c
Rank	2.48*	.11*	.148	5.16**	.11*	.347
Point assignment	2.21*	.09	.117	6.50**	.26*	.370
Forced choice	0.23	.06	-.004	2.88*	.12*	.182
Likert	0.83	.06	.053	0.92	.05	.090

^a Standard (Z) score adjusted according to the formula $Z = \bar{z} \sqrt{n(N-3)}$ where \bar{z} = average z for the sample, n = the number of subjects in the sample, and N = the number of pairs of scores in the correlations.
^b Proportion of intrasubject Pearson correlation coefficients that were significant, at least at the $p < .05$ level.
^c Average intrasubject Pearson correlation coefficient.
* $p < .05$. ** $p < .01$.

Intrasubject correlation coefficients were calculated between the scores provided by each of the four values measurement methods, the frequencies of value category responses to the perceptual task, and the weights obtained from within-subject regressions of responses to the decision task. This procedure provided a measure of the relation between scores on each of the values measures and the responses on the perceptual and decision tasks for each individual subject. To examine relationships between subjects, we cumulated intrasubject correlation coefficients through use of the Fisher r to z transformation (Wert, Neidt, & Ahmann, 1954). This procedure corrects somewhat for the skewness of the distribution of correlated variables. Results of this test are shown in Table 2. All results for ranks are reverse scaled to be consistent with other measures. The higher the score for a value, the more important the value to the individual.

Column 1 of Table 2 shows Z scores adjusted for significance testing for the values measures and the frequency scores on the perceptual task. Although these relations were not very strong, two measurement procedures were significant. Scores on both the ranking and point-assignment measures significantly related to perceptual responses at the $p < .05$ level. For these two measures, individuals tended to identify more often those values they had previously rated as more important in response to the ambiguous stimuli. This was not the case for the forced-choice and Likert measures.

Column 4 in Table 2 shows Z scores adjusted for significance testing for the values measures and the decision beta weights. Value orderings were found to relate significantly to decision weights for the rank, point-assignment, and the forced-choice measures at the $p < .05$ level or higher. Subjects emphasized or gave more weight to values on the decision task that they had previously rated more highly on these three measures. The Likert technique, however, did not result in a significant relation with decision weights.

In addition to the z transformation test, a proportions test was also used to evaluate the extent to which the number of significant (at the $p < .05$ level) within-subject correlations differed from chance findings. Table 2 displays these proportions in columns 2 and 5 for the perceptual and decision tasks, respectively. Inspection of Table 2 shows that the pattern is virtually identical to that produced by the Fisher's z transforma-

tion test. Using either test, the simple rank ordering of values appears to be the most highly related to task behavior of the four measurement approaches.

Average within-subject correlation coefficients are presented in columns 3 and 6 of Table 2 for the perceptual and decision tasks, respectively. Although none of these relationships explains more than 14% of the variance, these limited effect sizes are very much in keeping with our beliefs concerning the weak effect of values on behavior at any point in time.

Results relating the values measures under study here with Crandall's Social Interest Scale and the Pro-Protestant Ethic part of the Blood Protestant Ethic Scale are presented in Table 3. Chi-square statistics were calculated to determine if individuals who scored highly on the SIS were more likely to order *helping* as either their first or second priority. Between-subject analyses are inappropriate for analysis of ipsative data; therefore, chi-square tests were used instead. As can be seen from the table, the rank, point-assignment, and forced-choice measures all resulted in orderings that related to high and low scores on the SIS, as defined by a median split. Subjects who ordered helping either first or second on these three measures were more likely to score above the median on the SIS ($\phi = .345$, for the rank; .317, for the point-assignment; and .349, for the forced-choice measures). The Likert measure failed to reach significance, $\chi^2(1, N = 99) = 3.45, p = .06$.

Results for the Blood Protestant Ethic Scale were disappointing at two levels. First, past research has shown that the Pro-Protestant Ethic and the Non-Protestant Ethic subscales are generally negatively related (Blood, 1969). In the present study, however, the correlation coefficient for the two subscales was positive and highly significant ($r = .505, p < .001$). Second, chi-square analysis was used to ascertain the relation between value orderings on the four measurement approaches and the single value scale, the Pro-Protestant Ethic scale. For three of the measures tested here, the rank, point-assignment, and forced-choice methods, the ordering of achievement as either the first or second value versus the third or fourth was not related to high and low scores on the Pro-Protestant Ethic as defined by a median split (see Table 3). The Likert measure in this case outperformed the other measures tested. *Achievement*, as measured by this scale, was more likely to be ordered as less important for those subjects scoring below the median split on the Pro-Protes-

Table 3
Chi-Square Levels for Differences in Value Ranks as a Function of Social Interest, the Protestant Ethic, and Social Desirability for Different Measures of Values

Measure	Helping rank: high vs. low social interest ^a	Achievement rank: high vs. low protestant ethic ^b	First ranked values: high vs. low social desirability ^c
Rank	10.41*	0.093	1.49
Point assignment	7.62*	0.078	1.74
Forced choice	10.18*	1.23	1.32
Likert	3.45	6.62*	2.89

^a Comparison of the proportion who ranked helping either first or second for subjects scoring above vs. below the median on the Social Interest Scale; $df = 1$.

^b Comparison of the proportion who ranked achievement either first or second for subjects scoring above vs. below the median on the Pro-Protestant Ethic scale; $df = 1$.

^c Comparison of the pattern of the first ranked value for subjects scoring above vs. below the median on Social Desirability Scale; $df = 3$.

* $p < .01$.

tant Ethic, and ordered as more important for subjects scoring above the median ($\phi = .277$). One must view this result with some suspicion, given the overall poor performance of the Likert measure to this point. The measurement method of the Pro-Protestant Ethic subscale and the Likert measure are the same, suggesting that method variance may have some role in explaining these results. The SIS, on the other hand, is an adjective-choice method, which is unlike any other measurement approach used in this study.

A major concern with the measurement of values is that of a social desirability response set. The Marlowe-Crowne test for social desirability was used here to examine whether this was a problem for any of the measures used. Chi-square analysis was used to assess whether subjects were responding to social desirability cues rather than to their own values. Because the Likert measure allows for between-subject comparisons, this measure was also analyzed further.

Inspection of the point-assignment responses indicated that most subjects, if they did not have a full hierarchical ordering of values, had at least one value that was clearly preferred over the others, even if the individual expressed indifference between the remaining values. This first value appears to be important in determining general patterns of response to the various values measures. This suggests that one possible test of a measure's susceptibility to social desirability effects is to assess whether the pattern of first value choices was affected by scores on the Marlowe-Crowne test. The frequency with which each value was chosen first by the subjects was examined for both high and low scorers on the Marlowe-Crowne measure, as determined by a median split. If low scorers choose their first value to be achievement, helping, fairness, or honesty, more or less often than do high scorers, effects for social desirability biased responses are indicated. Results of these analyses for each of the four measures compared here are presented in Table 3. None of the measures showed any significant relation to scores on the Marlowe-Crowne scale.

Although each of the purely ipsative measures provides some control for socially desirable responses, this is not true of the Likert measure. Because we expected to see some relation between responses to this measure and the responses to the Marlowe-Crowne, we performed additional chi-square analyses, examining different patterns of responses. High and low scorers on the Marlowe-Crowne were compared for differences in the frequency with which they chose achievement, honesty, fairness, and helping as their first or second priority, versus third or fourth, in the belief that if one of these values was perceived as more socially desirable than the others, high scorers on the Marlowe-Crowne scale should choose that value first or second more frequently than should low scorers. None of these four additional tests had significant results, $\chi^2(1, N = 102) = .157, .354, .00$, and $.00$, for achievement, honesty, fairness, and helping, respectively.

Between-subject tests using the Likert measure were also performed. No direct comparisons can be made between the Likert and the ipsative measures on this test. Correlation coefficients calculated for both the total score on the Likert measure and for each of the four subscales with the Marlowe-Crowne score indicated small but significant relations ($r = .268, p = .003$, for the total Likert score; $r = .193, p = .026$, for the Achievement subscale; $r = .195, p = .025$, for the Helping subscale; $r = .293, p = .001$, for the Honesty subscale; and $r = .220, p = .013$, for the Fairness subscale). Although the size of these coefficients makes interpretation of their importance somewhat difficult, they do indicate that Likert responses were affected to some small degree by the concern for responding in a socially acceptable manner. To further examine the relation of the Likert measure to social desirability, a reliability analysis was performed for each subscale, resulting in extremely high interitem reliabilities (Cronbach's $\alpha = .959, .950, .946$, and $.952$, for achievement, fairness, honesty, and helping, respectively). Unfortunately, when four subscales were constructed using randomly selected items from the measure, the results were equally spectacular (Cronbach's $\alpha = .914, .947, .945$, and $.961$, for random subscales 1, 2, 3, and 4, respectively). This finding suggests that subjects were responding to all of the items in a similar way, possibly in part because they are equally socially desirable phenomena, despite the different values they represented.

Discussion

Partial or full support was found here for four of the five hypotheses of this study. Hypothesis 1 received partial support. Both rank and point-assignment scores related to responses on the perceptual task, whereas Likert and forced-choice scores did not. Hypothesis 2 was fully supported. Subjects emphasized values in decision making in relation to their importance as indicated on the rank, point-assignment, and forced-choice measures, whereas Likert responses were not related to decision weights. Hypothesis 3 received partial support. Using between-subject statistics, some slight influence of social desirability response bias was found for the Likert measure. However, no effects were found using the chi-square analysis necessary to compare the Likert with the ipsative values measures. Hypothesis 4 was not supported in that results were the reverse of those predicted. Subjects' responses to the Pro-Protestant Ethic sub-

scale were not related to responses on the rank, point-assignment, or forced-choice measures, but were related to Likert responses. This finding may have been caused in part by common method variance. Hypothesis 5 was supported by the finding that the rank, point-assignment, and forced-choice responses were related to scores on the SIS, whereas Likert responses were not.

Although validities for all of the measures are somewhat low, over multiple criteria, the rank-order measure performed the closest to predictions in its relation with cognitive hierarchical orderings, decision behavior, established measures, and control of social desirability, whereas the Likert measure showed little relation to performance on the perceptual and decision tasks, and also correlated with the Marlowe-Crowne measure of social desirability response bias. These findings tend to confirm our doubts about the ability of the Likert measure to control for socially desirable responses, and to raise the question of whether other scales using this method of measuring many values or a single value are not also subject to this problem. Not only did the rank measure give the overall best performance against the criteria, it is the simplest and least time consuming for both the subject and the researcher. The point-assignment method also performed very well overall, whereas the forced-choice measure did less well, possibly because its complexity was less congruent with the simple perceptual and decision tasks presented here. These findings unfortunately leave us with the problem of the analytic limitations of ipsative measures. An important challenge for future research is to rigorously question the hierarchical model of values and the issue of whether individuals maintain stable preference orderings over time. This conceptualization is one primary rationale for the ipsative measurement model. If it does not hold up under scrutiny, other means of controlling response bias should be devised to eliminate the need for ipsative measurement.

Establishment of a nomological network for values measures is an interactive process, which requires that both measurement and theory be investigated simultaneously (McGrath, 1982); thus, we believe that the research reported here also provides some positive support for work values theory. Findings indicate that an individual's values are related to his or her cognitive organization. Subjects were given no reason to believe that their primary value would appear any more often than the other values in the nonsense words flashed on the screen in the perceptual task, nor were there any influences that might suggest to them that they should somehow be consistent with questionnaires they had filled out days earlier. In effect, this was a very stringent criterion inasmuch as interpretation was at a purely subconscious level. Despite these conditions, subjects tended to "recognize" words as belonging to a value category with frequencies corresponding to their value hierarchies. This finding gives some support for the notion that values are indeed hierarchically organized in memory, and strongly suggests that people will find opportunities, within the context of their duties, to apply their dominant value in uncertain situations. That is, these people are likely to perceive the implications for that value in situations that may be interpreted in multiple ways. An organization that desires to promote a cooperative climate, for example, should therefore prefer to employ individuals who could

formulate opportunities for cooperative behavior under conditions in which the appropriate behavior is not predetermined.

Values were also found to act as a guide or standard for decision making. In addition to its theoretical significance, this finding has profound implications for organizations that desire decisions to be reflective of particular patterns of values. Clearly, selection of individuals who hold a particular dominant value would, over the long term, influence the overall value orientation of company decisions. Of course, this study may well have magnified the effect values have on decision making (subjects had essentially no other information beyond the value information concerning criteria for making a decision). In more complex, less compacted environments, the relation of values to choices should be a weak longitudinal one, given that people choose alternatives for reasons other than their system of values.

As with any study that administers many tests during a short period of time, one must consider the problem of consistency effects. Subjects may actively attempt to be consistent across all questionnaires and other activities. In this study, the order of the various tasks was chosen to separate in time as far as was possible the items most subject to this effect (ranks and point assignments). Other tasks were designed such that a subject should either have no reason to be consistent (the perceptual task) or should have difficulty in attempting to consciously control consistency (the forced-choice measure and the decision task). An additional problem is that the sample used here, undergraduate students, is not ideal; however, the length of the comparative procedure and the need for research into cognitive organization precluded use of a working sample. Research on a working sample would certainly help to support the results found here.

This study suggests many paths for future research. Perhaps the most important of these is to study the impact of values on individual processes. Further exploration of the relation between values, decision making, and behavior is needed both in the field and under more controlled conditions over longer time periods. This should include commonly used dependent variables such as satisfaction and commitment, as well as the central variables considered here. The use of multiple methods for purposes of triangulation is particularly advocated, given the moderately good performance of three of the values measures examined here. Finally, the initial establishment of evidence relating values to cognitive organization indicates a need for more research concerning the cognitive processes whereby values influence choice behavior.

References

- Allport, G. W., Vernon, P. E., & Lindzey, G. (1960). *Allport-Vernon-Lindzey study of values* (3rd ed.). Boston: Houghton Mifflin.
- Anastasi, A. (1982). *Psychological testing* (5th ed.). New York: Macmillan.
- Blood, M. R. (1969). Work values and job satisfaction. *Journal of Applied Psychology*, 53, 456-459.
- Cherrington, D. J. (1980). *The work ethic: Working values and values that work*. New York: AMACOM.
- Cornelius, E. T., Ullman, J. C., Meglino, B. M., Czajka, J., & McNeely, B. (1985, November). *A new approach to the study of worker values and some preliminary results*. Paper presented at the meeting of the Southern Management Association, Orlando, FL.

- Cottrell, N. B., Wack, D. L., Sekerak, G. J., & Rittle, R. H. (1968). Social facilitation of dominant responses by the presence of an audience and the mere presence of others. *Journal of Personality and Social Psychology*, 9, 245-250.
- Crandall, J. E. (1975). A scale for social interest. *Journal of Individual Psychology*, 31, 187-195.
- Crandall, J. E. (1977). Further validation of the social interest scale: Peer ratings and interpersonal attraction. *Journal of Clinical Psychology*, 23, 140-142.
- Crandall, J. E., & Harris, M. D. (1976). Social interest, cooperation, and altruism. *Journal of Individual Psychology*, 32, 50-54.
- Crowne, D. P., & Marlowe, D. (1964). *The approval motive: Studies in evaluative dependence*. New York: Wiley.
- Deal, T. E., & Kennedy, A. A. (1982). *Corporate cultures: The rites and rituals of corporate life*. Reading, MA: Addison-Wesley.
- England, G. W. (1967). Organizational goals and expected behavior of American managers. *Academy of Management Journal*, 10, 107-117.
- England, G. W. (1975). *The manager and his values: An international perspective from the United States, Japan, Korea, India, and Australia*. Cambridge, MA: Ballinger.
- Epstein, S. (1979). The stability of behavior: 1. On predicting most of the people much of the time. *Journal of Personality and Social Psychology*, 37, 1097-1126.
- Epstein, S. (1980). The stability of behavior: 2. Implications for psychological research. *American Psychologist*, 35, 790-806.
- Fallding, H. (1965). A proposal for the empirical study of values. *American Sociological Review*, 30, 223-233.
- Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, 51, 327-358.
- Henchy, T., & Glass, D. C. (1968). Evaluation apprehension and the social facilitation of dominant and subordinate responses. *Journal of Personality and Social Psychology*, 10, 446-454.
- Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin*, 74, 167-184.
- Howard, A., Shudo, K., & Umeshima, M. (1983). Motivation and values among Japanese and American managers. *Personnel Psychology*, 36, 883-898.
- Kluckhohn, C. (1951). Values and value-orientation in the theory of action. In T. Parsons, & E. Shils (Eds.), *Towards a general theory of action* (pp. 388-433). Cambridge, MA: Harvard University Press.
- Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping*. New York: Springer.
- Locke, E. A. (1975). Personnel attitudes and motivation. *Annual Review of Psychology*, 26, 457-480.
- Locke, E. A. (1976). The nature and consequences of job satisfaction. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 1297-1349). Chicago: Rand-McNally.
- Locke, E. A. (1982). *A new look at motivation: Theory V* (Tech. Rep. GS-12). College Park: University of Maryland, College of Business and Management.
- Locke, E. A., & Henne, D. (1986). Work motivation theories. In C. L. Cooper & I. Robertson (Eds.), *International review of industrial and organizational psychology* (pp. 1-35). Chichester, England: Wiley.
- McGrath, J. E. (1982). Dilemmatics: The study of research choices and dilemmas. In J. E. McGrath, J. Martin, & R. A. Kulka (Eds.), *Judgment calls in research* (pp. 69-102). Beverly Hills, CA: Sage.
- Miner, J. B. (1977). *Motivation to manage*. Atlanta, GA: Organizational Measurement Systems Press.
- Ouchi, W. G. (1981). *Theory Z*. New York: Avon.
- Peters, T., & Waterman, R., Jr. (1982). *In search of excellence*. New York: Harper & Row.
- Postman, L., Bruner, J. S., & McGinnies, E. (1948). Personal values as selective factors in perception. *Journal of Abnormal and Social Psychology*, 43, 142-154.
- Rokeach, M. (1973). *The nature of human values*. New York: Free Press.
- Schein, E. H. (1985). *Organizational culture and leadership*. San Francisco: Jossey-Bass.
- Spence, J. T. (1985). Achievement American style: The rewards and costs of individualism. *American Psychologist*, 40, 1285-1295.
- Staw, B. M. (1984). Organizational behavior: A review and reformulation of the field's outcome variables. *Annual Review of Psychology*, 35, 627-666.
- Wert, J. E., Neidt, C. O., & Ahmann, J. S. (1954). *Statistical methods in educational and psychological research*. New York: Appleton-Century-Crofts.
- Williams, R. M., Jr. (1968). The concept of values. In D. Sills (Ed.), *International encyclopedia of the social sciences* (pp. 283-287). New York: Macmillan.
- Yankelovich, D. (1979). Work, values, and the new breed. In C. Kerr & J. M. Rosow (Eds.), *Work in America: The decade ahead* (pp. 3-26). New York: Van Nostrand Reinhold.
- Zajonc, R. B., & Nieuwenhuysse, B. (1964). Relationship between word frequency and recognition: Perceptual process or response bias. *Journal of Experimental Psychology*, 67, 276-285.
- Zajonc, R. B., & Sales, S. M. (1966). Social facilitation of dominant and subordinate responses. *Journal of Experimental Social Psychology*, 2, 160-168.

Received May 6, 1986

Revision received April 6, 1987

Accepted January 26, 1987 ■

Organizational Determinants of Leader Behavior and Authority

Tove H. Hammer and Jay M. Turk
New York State School of Industrial and Labor Relations, Cornell University

We examined the effects of technology, the organization's dependence on a section or unit to attain its mission, labor union strength in the workplace, and managerial pressure for close, strict supervision on four dimensions of leader behavior and leader authority to direct work. Data on leadership came from 160 first-line supervisors in 12 sections of a manufacturing plant. Regression analyses showed that technology, union strength, and management pressure contributed significantly to responsive leader behavior. Supervisors in long-linked technologies spent more time on work-group maintenance; supervisors in intensive technologies spent more time on developing and maintaining links with other administrative units; supervisors in heavily unionized sections interacted with subordinates more according to written rules and regulations; and supervisors whose upper level managers favored close, strict supervision responded by pushing subordinates to work in a punitive manner. Technology dominated the prediction equation for authority. Supervisors in intensive technologies had more authority than supervisors in long-linked technologies.

It has long been acknowledged that what first-line supervisors can do to shape work-group performance and organizational effectiveness is circumscribed by factors outside their control. Examples are trade union presence (Homans, 1965), the nature of work carried out by subordinates (House & Mitchell, 1974), characteristics of the work force (Filley, House, & Kerr, 1976), government regulations of personnel policies (Hammer, 1979), and demands from superiors, subordinates, and peers (Fleishman, Harris, & Burt, 1955; Lowin & Craig, 1968; Pfeffer & Salancik, 1978; Rosen, 1969).

Except for studies examining the effects of subordinates' and superiors' expectations and performance on supervisors' behaviors, little empirical information exists on the external determinants of leadership activities. But recent theoretical models have incorporated organizational constraints by separating those aspects of the leader role that originate in the leader from those caused by outside forces. We use three models as a theoretical foundation for a study of the effects of organizational characteristics on first-line supervisors' leadership behaviors and authority: the resource-dependence model of organizational control (Pfeffer & Salancik, 1978), the multiple influence leadership model (Hunt & Osborn, 1982), and Stewart's demand-constraints-choices model of managerial jobs (Stewart,

1982). We will briefly describe these models and their usefulness for this research.

Pfeffer and Salancik (1978) identified three managerial (or leadership) roles: the *symbolic*, in which the manager is a symbol of the organization and its success or failure; the *responsive*, in which the manager processes information from the environment and responds to the demands and constraints that confront the organization; and the *discretionary*, in which the leader acts to change the environment to the organization's benefit. Our interest is in the responsive role, in which the leader's activities are structured and shaped by outside events. A responsive leader has choices, but they are limited to decisions about which demands to meet and which to ignore or reject.

Hunt and Osborn (1982) separated leader behaviors into two broad categories: (a) discretionary activities that leaders engage in on their own initiative and (b) responses to demands from people near, or at, their own organizational level (so-called *lateral* behavior).

Stewart (1982) argued that managers (beyond the first-line supervisory level) have considerable choice about what gets done and how it gets done. But choices are restricted by demands—tasks that must be done by the jobholder only—and constraints—factors internal and external to the organization that limit what the manager can do. Stewart's thesis is that managerial jobs are much more inherently flexible than is frequently assumed. Her research identifies broad classes of demands and constraints used to describe differences between managerial jobs, and differences between managers in similar jobs. Of interest to us in Stewart's model are the organizational demands and constraints that dictate what must be done (in the case of demands) and what ought to be done (in the case of constraints).

Neither the resource-dependence nor the multiple influence model sheds much light on specific determinants of responsive leadership or on the identification of categories of responsive behavior. Pfeffer and Salancik (1978) did not develop the leadership component of their model into a set of hypotheses. The multiple influence model was developed to predict work-unit (subordinates) performance from leader behavior. Hunt and

This study is based, in part, on discussions between the senior author and Ned Rosen about the impact of technology on leader behaviors. They collected the data for the study as part of a joint project on leadership, and Ned Rosen contributed substantially to the instrumentation phase of the research. An earlier, shortened version of the article was presented at the 45th Annual Meeting of the Academy of Management, San Diego, August 1985.

The authors would like to thank Steven Currall, Robert Doherty, and Ned Rosen for their helpful comments on earlier drafts, and Ivor Francis and Leonard Stefanski for their statistical advice.

Jay Turk is now at Data General in Boston, Massachusetts.

Correspondence concerning this article should be addressed to Tove Hammer, New York State School of Industrial and Labor Relations, Cornell University, Ithaca, New York 14851-0952.

Osborn (1982) concentrated their theoretical efforts on the definition of discretionary leadership because, they argued, it has more potential than does responsive leadership for changing subordinates' attitudes and performance.

In this study, we use the classification of leadership identified in the resource-dependence and multiple influence models and select our set of predictors in part from them and in part from Stewart's (1982) work. No new model of leadership is developed. Instead, we focus on the first-line supervisor's responsive role and test a series of hypotheses about relations between organizational characteristics and dimensions of leader behavior and authority.

Organizational Constraints

A large number of variables in a supervisor's environment can influence how time is spent. In this study, we are interested in examining the possible effects of individual organizational-level constraints over which supervisors have no control. The following constraints are included: technology, the organization's dependence on a work unit or section (the unit's importance to the organization's overall goal or mission), the strength of a labor union presence, and a management philosophy that favors tight or close supervision.

Technology is included for two reasons. First, it determines the nature of work available for subordinates and also to some extent, the nature of the subordinates themselves. Both job tasks and work-force characteristics have been found to constrain leader behavior and effectiveness (Filley et al., 1976; House & Mitchell, 1974; Kerr & Jermier, 1978; Lowin & Craig, 1968). Second, technology is a central variable in both the multiple influence model (Hunt & Osborn, 1982) and in Stewart's (1982) research.

The organization's dependence on a work unit or section comes from the resource-dependence model, in which it is considered a source of intraorganizational power for unit or section management. But although centrality to the organization's mission can be a management power base, it requires work performance standards that will be passed on as demands from section leaders to first-line supervisors. There is research evidence for the constraining nature of managerial pressure for performance on supervisors' behavior (Pfeffer & Salancik, 1975).

The presence of a trade union means that formal interactions between management and labor are regulated by contract stipulations and, at times, by practices developed over time (past practice). A union contract specifies how workers should be treated in the labor process, the traditional domain of the first-line supervisor. The complaint that supervisors have lost both authority to direct the work force and freedom to behave as they please to the union, is of old standing (see Homans, 1965). Union agreements are also on Stewart's (1982) list of constraints.

The last constraint variable is the organization's philosophy about the management of labor, less tangible than technological processes and trade union presence, but a supposedly important factor in determining leadership behaviors (Fleishmann et al., 1955). It is included in this study because it represents pressure on supervisors that is generated by upper management's definitions of appropriate leadership styles, and because there is a prevailing assumption that it is an important constraint on

the supervisor, leading to an undue amount of close supervision.

Leader Behavior and Authority

The research literature contains a number of conceptual and operational definitions of leader behavior that vary in specificity. The most widely known and used definitions, such as the Ohio State leadership questionnaires and the Michigan leadership scales (see Stogdill, 1974), measure broad categories of behavior. More specific definitions that capture a wide variety of behaviors, such as the Yukl and Nemeroff (1979) leader behavior scales, do not distinguish between discretionary and responsive leadership. The Hunt and Osborn (1982) definition of responsive (or lateral) leadership has three dimensions: interaction with members of other work units or the development of interunit networks, pressure for action from others, and adaptation to such pressure.

Given the limited amount of data available on responsive leadership, two approaches were used to identify leader behavior dimensions. First, we identified among existing leadership definitions those dimensions that could be considered responsive behavior (e.g., Hunt & Osborn's, 1982, "networking"; Yukl & Nemeroff's, 1979, "conflict management"; the traditional "production or performance emphasis"). At times, it was difficult to separate narrowly defined leadership dimensions into discretionary and responsive behaviors. Some behaviors, such as performance emphasis, could be in either category, depending on the supervisor's circumstances. We then identified those responses that would be required of supervisors in the presence or absence of our independent variables. This determined the choice of leader behaviors. The ones selected overlapped to a large extent with dimensions available in the leadership literature.

With leader authority, we mean the power to direct work vested in first-line supervisory positions by the organization and recognized as legitimate by organization members. We do not include here the ability to influence subordinates. How supervisors use the power granted them to influence others is a different matter.

Hypotheses

More precise definitions of the independent variables are necessary before the hypotheses linking them to responsive leadership become obvious. We begin with technology, by which we mean the extent of mechanization and automation of the production process (Hall, 1982). Specifically, we use Thompson's (1967) model, which classifies organizational units (sections, divisions, departments) into three technology types: *long-linked*, *mediating*, and *intensive*.

Long-linked technology is a set of linked, interdependent tasks or operations, analogous to a mass-production assembly line. The work process is rationalized, routinized, simplified, short cycled, and fast paced. Workers are usually semiskilled or unskilled. A mediating technology involves activities that link clients or customers who are, or who wish to become, connected. Examples of organizations serving mediating or linking functions are banks or insurance agencies. Tasks are more varied, and require some training or prior preparation of workers.

In an intensive technology, tasks are complex and nonroutinized. A variety of techniques and work methods may be used at the point of production depending on the nature of the job. A research laboratory or a trouble-shooting maintenance and repair shop are examples. Specialized skills and professional knowledge are required of workers in intensive technologies, who are, therefore, skilled labor.

The routine, simplified tasks and the low-skilled labor in the long-linked, or mass-production, technologies have two implications for supervisors. First, the nature of the tasks themselves will structure the work. There should be relatively little need for supervisors to instruct, demonstrate, or explain contingencies between work behavior and performance outcomes. The classic *initiation of structure* will be redundant (House & Mitchell, 1974, and Kerr & Jermier, 1978, have more detailed theoretical explanations from the path-goal model on this point). In addition, low-skilled labor is often entry-level labor with high turnover rates. In long-linked technologies, there is also lateral movement of workers through frequent job changes (Vardi & Hammer, 1977). Labor-force instability and low-level labor should mean that the supervisor needs to spend time on work-group maintenance, such as integrating new people into work units, explaining rules and regulations, and resolving differences between new and old workers (Stewart, 1982, presents a very similar argument).

In contrast, work in an intensive technology, in which skilled and professional workers do more complex jobs that are linked to the activities of other departments or work units, should not require group maintenance from the leader because work groups are more stable (Vardi & Hammer, 1977) and workers are more internally motivated and committed to their work (Filley et al., 1976). Instead, the nature of work will require of the supervisor activities such as attending meetings, being in contact with interlocking departments, negotiating the relation between work units—a form of *linking-pin* behavior or *boundary spanning*, or what Hunt and Osborn (1982) call *networking*, which tie or link performance in his or her unit to that of other units.

Hypothesis 1: Working in a long-linked technology will be positively related to the amount of time supervisors spend on work-group maintenance. In particular, supervisors in long-linked technologies will spend more time on work-group maintenance than will supervisors in intensive technologies.

Hypothesis 2: Working in an intensive technology will be positively related to the amount of time supervisors spend on boundary spanning or network development. In particular, supervisors in intensive technologies will spend more time on networks than will supervisors in a long-linked technology.

The authority supervisors have to direct work, to make decisions about work assignments, procedures, and performance evaluations, should also vary with technology. As the production process becomes more unpredictable (less determined by technological specifications), more discretion is required by both workers and managers to translate production goals into action. Following Pfeffer and Salancik's (1978) arguments, the organization is more dependent on its first-line supervisors to operationally define the jobs to be done (and thus get them accomplished) in intensive technologies than in long-linked technologies, and is therefore more likely to grant them more authority.

Hypothesis 3: Supervisors working in an intensive technology will have more authority to direct work than will supervisors working in a long-linked technology.

The more the organization depends on a unit or section for the accomplishment of organizational goals, the more powerful the unit is, relative to other parts of the organization, and the more resources it can demand (Pfeffer & Salancik, 1978). This argument is reinforced by Jaques (1976), who found that the more central a unit is to the organization's core, the more power accrues to its management and the longer is the management's *timespan of discretion* (the period of time that can elapse before the management of the unit is subject to scrutiny from above). But there are constraints on leaders in central units. The costs of making mistakes and the visibility of mistakes are high, and there is pressure for performance, which will be passed on to first-line supervisors. Supervisors should react by putting pressure on subordinates to fulfill performance expectations.

Hypothesis 4: The organization's dependence on a section or work unit will be positively related to supervisors' pushing subordinates to perform.

Because mistakes are visible and the costs of making them are high in core units, we expect that the authority to direct the production process will be centralized at the unit-management level. This means that first-line supervisors should have less discretion in the exercise of *their* work and less authority to make labor process decisions.

Hypothesis 5: The organization's dependence on a section or work unit will be negatively related to first-line supervisor's authority to direct work at the point of production.

Working with a unionized labor force means that supervisors have to follow standard rules and regulations in interactions with subordinates. The union contract spells out labor's rights and management's obligations. A strong union presence should require that supervisors consult the contract and generally follow whatever prescriptions and guidelines are set forth for the treatment of workers.

Hypothesis 6: Union strength will be positively related to supervisors' attending to formal rules and guidelines in the treatment of workers. In particular, the stronger the union presence, the more supervisors will go by the book.

A strong union could diminish supervisors' overall authority. However, by helping to codify the supervisor's legitimate role as part of management, the union might actually reinforce it. Because the union's effect on authority can be equivocal, no directional hypothesis is stated.

The reason for arguing that a management philosophy about appropriate leadership styles will constrain supervisory behavior has been presented. The formal hypothesis states the following.

Hypothesis 7: Working for a management whose philosophy of leadership demands tight or close supervision will be positively related to the extent to which supervisors supervise closely.

Method

Research Setting

The study was done in a large pharmaceutical plant located in the eastern United States. The plant, which belongs to a major manufactur-

ing company, has about 3,000 employees and includes both production and service sections. The bulk of production is done in large chemical and pharmaceutical sections, supported and serviced by production planning, repair and maintenance, industrial engineering, packaging, printing, shipping, and small research sections. Most functions of quality control, large-scale research and development, and personnel management are done by central units that serve the entire company.

The study included 12 sections that constitute the plant itself. Sections are divided into different functional departments, within which work is divided between various work groups directed by first-line supervisors. Plant employees are represented by the International Chemical Workers Union and are covered by union contract. Because the plant is not a union shop, union membership is voluntary and the percentage of unionized employees differs across sections.

Samples and Procedures

Data came from two samples within the plant and from plant archives. The primary sample comprised 160 first-line supervisors (144 men, 16 women) from the 12 sections included in the study. There were a total of 167 supervisors in the whole plant; of these, 7 did not participate due to illness, vacations, and other absences during data collection. Supervisors completed questionnaires designed for this study on company time as part of a 2-year organizational development and training program conducted for the plant by the senior author and a colleague (see author note). The sample had a mean age of 49 years, a mean education level of 13 years, and a mean of 24 years of plant tenure.

The second sample comprised three section-level managers from different parts of the plant with special knowledge of plant-wide organizational characteristics. They provided data about a series of section-level variables through rating forms and interviews. Information about the union came from union and company archives.

The questionnaires developed for the supervisors were based on interviews conducted by the senior researcher with representatives from section- and department-level management, and were designed to include information of concern to this population of supervisors. All data were confidential and were fed back to groups of supervisors and their managers as part of the training program in the form of section-level aggregates.

Measures

Organizational characteristics. The classification of sections by technology was done by the senior author on the basis of information obtained through interviews with the three experts who knew the technological characteristics of all plant units, and through observations by the researcher from plant tours and visits. Four sections were classified as having a long-linked technology: two batch mass-production units, a bottling and packaging unit, and a plant sanitation and cleaning section. (The cleaning unit could be considered a service unit, but it did not satisfy Thompson's, 1967, criteria for a mediating technology because it did not serve as a linkage between other units. Work was routine, tasks simple, and the work process rationalized.) A total of 73 supervisors worked in a long-linked technology. Two sections, production planning and control, and material handling (shipping and receiving), with 18 supervisors, were classified as mediating technologies. Six sections with 69 supervisors were identified as having an intensive technology: two repair and maintenance sections (craft workers), one industrial engineering unit, two research laboratories, and one print shop doing small-scale, custom jobs.

The classification process and results were discussed with a colleague of the senior author, who was familiar with the plant from earlier studies of job mobility (Vardi & Hammer, 1977).

Union strength was measured with a union density index, the percentage of workers in a section who belonged to the union. Percentage organized is a common operational definition of union strength in the

collective bargaining and labor economics fields, but it has been criticized for deficiencies when it is used on the aggregate level (see Kochan & Block, 1977). In this case, union strength was conceptually defined as union density, so percentage organized is therefore the appropriate measure. Percentage organized ranged from 0 to 82, with a median value of 60.

The organization's dependence on the section (in this case, the plant's dependence) and managerial philosophy about close or tight supervision were measured with ratings from the three plant experts. Included in the rating task was the measurement of several other organizational characteristics used to test assumptions made in this study about processes leading from technology and a section's centrality to the core to leader behavior. Each rater was asked to rank-order the 12 sections on 10 characteristics. Sections were listed on special rating forms in alphabetical order, and the rank-order procedure and the definition of the characteristics were described in detail on each page of the rating form (one variable per page).

Organizational dependence on section was defined as "the significance of each unit's mission and contribution to the total organization's achievements," and the rank-order task was described as involving the evaluation of "the unit whose mission is such that the larger organization is most (least) dependent on it . . . to what extent does X's (the plant) survival depend on the function performed by each unit?"

Management philosophy was called "philosophical climate" on the rating form, and was described as follows:

The extent to which the supervisors of a given unit are intentionally pressed by higher managers to adopt or utilize a given style of leadership in running their unit. Thus, a unit might be under pressure from middle to higher management to become either more or less autocratic in decision making, to supervise rank and file either more or less closely and strictly, to be more or less human relations conscious, and so on . . . rank the 12 units on the extent to which they are under deliberate management pressure to adopt a supervisory-leadership style characterized by close, strict supervision.

In addition, the three experts ranked the 12 sections on the following characteristics: labor force stability, presence or absence of a Protestant work-ethic culture among workers, costs of making mistakes (such as failing to meet objectives), visibility of mistakes, financial/budgetary significance of sections, task complexity, interdependence of work roles in section (coordination requirements), and work-load volume.

Rank-order data were collected because the senior author and Ned Rosen (see author note) believed they would be more valid measures of the organizational characteristics examined in the study than would data from rating scales. The raters had said during initial interviews that they would be able to discriminate between sections by ranking, and our aim was to capture these differences.

Interrater reliability estimates were calculated for all dimensions. The organizational dependence index had a reliability estimate of .86 (based on the average interrater rank-order correlation), and the managerial philosophy index had a reliability estimate of .81.

Leader behavior. The four leader-behavior dimensions were measured with two different instruments. A scale developed specifically to measure responsive leadership was used to assess work-group maintenance and network activities. Supervisors were asked to consider the duties that their present position required and to state how much time was spent performing a set of 24 tasks or activities. A 9-point verbally anchored scale was used, with response alternatives ranging from *very little time* (1) to *a great deal of time* (9). The requirement scale also included items concerning administrative activities, long-range planning, and direct supervision. Factor analysis produced three distinct factors: work-group maintenance, network development, and direct supervision. Scale items, factor loadings, and reliability estimates for the work-group maintenance and network activities are listed in the Appendix.

Performance emphasis and go-by-the-book activities were measured

with items from a supervisor-behavior index, developed to assess dimensions of leadership related to path-goal theory (Hammer, 1974; Hammer & Dachler, 1975). The scale was designed for use by subordinates and measured four dimensions of behavior, of which performance emphasis through the setting of performance-outcome contingencies was one. Scale items were rewritten to allow supervisors to describe their own behavior, and items were added about consulting company regulations and the union contract in the treatment of subordinates. Respondents were asked to state how frequently or how often they engaged in 38 activities, using a 7-point verbally anchored scale, with response alternatives ranging from *never* (1) to *all the time* (7). Factor analysis produced three clearly defined factors, one of discretionary leadership (work facilitation) and the others that included performance emphasis items and go-by-the-book behaviors (see Appendix).

Leader authority. Overall authority to direct work was measured with a 15-item scale that asked supervisors to state, on a 5-point scale, what level of authority they had over different areas. The scale is reproduced in the Appendix.

Demographic variables. Information was also collected on supervisors' age, years of formal education, plant tenure, and job tenure.

Results

The hypotheses were tested with hierarchical multiple regression analysis. Reasons for the order of entry of variables will be presented with the regression results. The same prediction equation was used for each dependent variable, and individual hypotheses were tested by examining the contribution of the appropriate independent variable to R^2 . Two predictors were assessed with rank-order data. These were normalized before any analyses were done and were transformed to 9-point scales. The shapes and properties of the marginal distribution of all variables included in the study were checked for possible deviations from normality. All variables had normal distributions; measures of skewness and kurtosis were less than ± 1 . Before regression results are presented, correlations between sets of organizational characteristics are discussed to allow examination of assumptions made about the effects of organizational dependence on the work process, the labor force, and the importance of organizational units. *T* tests were used to examine differences between long-linked and intensive technologies.

Relations Between Organizational Characteristics

For units central to the organizational core, it was assumed that production and budgetary mistakes would be costly and visible. Table 1 shows the correlations that support this assumption.

In addition, strong negative correlations were found between percentage unionized and labor-force characteristics, a pattern repeated for managerial philosophy about strict supervision. This means that the large production sections, which had mostly unskilled and semiskilled labor, were the most heavily unionized. These were the most important units for the organization's overall mission (production). They had the heaviest work-load requirements, and they experienced managements who favored close and strict supervision.

We argued that specific technologies would require specific leader responses because technology would determine the work processes and, in part, the skill and stability level of the work force. *T* tests for differences between long-linked and intensive technologies showed that the labor force was rated as more sta-

Table 1

Rank-Order Correlations Between Organization Dependence (OD), Union Density (UD), Management Pressure (MP), and Section-Level Organizational Characteristics (N = 12)

Characteristic	Reliability	OD	UD	MP
Labor stability	.78	-.03	-.80**	-.63*
Labor work ethic	.85	.26	-.52	-.60*
Costs of mistakes	.95	.83**	.04	.16
Visibility of mistakes	.53	.46	.16	.45
% of budget allocated to unit	.80	.36	-.09	.32
Work-role interdependence	.85	.61*	.19	.44
Work volume	.81	.41	.43	.81**
Task complexity	.84	.52	.04	.19

* $p < .05$. ** $p < .01$.

ble in the intensive technology units ($M = 6.16$ vs. $M = 3.25$; $t = 2.63$, $p < .05$). There was no significant difference between technologies with respect to task complexity, but work-load volume (the amount of work) was higher in the long-linked technology ($M = 6.75$ vs. $M = 3.83$, $t = 2.03$, *ns*).

Costs of making mistakes and the visibility of those mistakes were significantly higher in the long-linked than in the intensive technology ($M = 6.75$ vs. $M = 3.83$; $t = 2.63$, $p < .05$, and $M = 7.0$ vs. $M = 3.0$; $t = 2.77$, $p < .05$, respectively), which was expected because the long-linked technology units were those central to the organization's core.

The data on work process and labor-force characteristics for the mediating technology units fall between the data from long-linked and intensive technologies. Because the sample of sections is so small ($N = 12$), one-way analyses of variance were not significant.

Intercorrelations between all variables included in the regression analyses, except technology, are presented in Table 2. For these and subsequent analyses, the sample size was reduced to 156 because a pattern of extreme, nonvariable values on the behavior and authority scale items for 4 subjects made us doubt the validity of the data.

Regression Analyses

Each of the leader behavior dimensions and leader authority were regressed on the same set of predictors. The equation included the following variables, listed in order of entry: demographic characteristics (age, education), organization's dependence on section, technology (as a dummy variable with mediating technology as the reference category), union density, and managerial pressure for close supervision.

A hierarchical regression model was used because there was some joint variance among the predictors, and because the order of entry could be determined by the degree to which the predictors (in theory) had an immediate effect on the supervisor's work. There were two exceptions. One, demographic characteristics were entered first as control variables. Organizational tenure was not included as a predictor because it correlated .67 ($p < .01$) with age, and we wished to minimize the number of predictors with $n = 156$. Two, there was no theoretical reason why centrality to the core would be a more immediate force on the supervisor than was technology, but technology

Table 2
Intercorrelations of Organizational Characteristics, Leader Behavior, and Authority, Age, and Education

	1	2	3	4	5	6	7	8	9	10
1. Age	—	-.35*	-.20	.14	-.01	.25*	.22	-.07	-.02	.21
2. Education		—	.04	-.47*	-.33*	-.08	-.31*	-.21	.10	.08
3. Organizational dependence on unit			—	-.02	.18	-.23	.09	.11	.15	.07
4. Union density				—	.55**	.06	.33*	.22	-.22	-.02
5. Management pressure					—	.11	.11	.12	-.19	-.10
6. Performance emphasis						—	.40*	.17	.13	.18
7. Go-by-the-book							—	.36*	.02	.13
8. Work-group maintenance								—	.19	.04
9. Network activity									—	.34*
10. Overall authority										—

Note. $N = 145$ –156.

* $p < .01$.

so dominated all aspects of the organizational environment that centrality was entered first to maximize its contribution. Union density and management philosophy were thought to be less immediate factors, and were therefore entered in the last two steps.

Hypotheses 1–3 identified the relations between technology and work-group maintenance, network behaviors, and authority. Table 3 contains the results of the regression analyses. The sample size for the regression was reduced to 145, due to missing data.

Technology was significantly related to both responsive leadership and authority. Work-group maintenance was more prevalent among supervisors in long-linked technologies than in in-

tensive technologies, which supports Hypothesis 1. Supervisors were more involved in network and linking-pin activities in intensive technology units than in long-linked technologies, which supports Hypothesis 2. Supervisors in long-linked technologies reported having less authority than their counterparts in intensive technologies, supporting Hypothesis 3.

Of the other variables in the equation, years of formal education and managerial pressure for close supervision added significantly to the variance explained in group maintenance behavior. Supervisors with less education and less pressure from management spent more time responding to intragroup conflicts and subordinates' counseling and training needs. The full

Table 3
Hierarchical Multiple Regression Results With Predictors of Supervisor Behavior and Authority

Predictor	Cum R^2	F	B	Cum R^2	F	β	Cum R^2	F	B
	Work-group maintenance			Network activities			Performance emphasis		
Age	.01	1.21	-.09	.00	<1	.01	.07	10.46**	.25
Education	.07	8.94**	-.26	.02	2.19	.13	.07	<1	.01
Organization dependence on section	.07	<1	.05	.03	1.79	-.12	.10	5.45*	-.19
Technology									
Long-linked	.10	4.05*	.19	.16	21.23**	-.42	.12	3.74	-.18
Intensive	.14	6.97*	-.39	.17	1.74	.19	.12	<1	.01
Union density	.15	2.15	.13	.19	4.21*	-.18	.13	<1	.07
Management pressure	.18	4.36*	-.22	.19	<1	.01	.17	7.62*	.29
R^2 adjusted	.14			.15			.13		
F value for the equation	4.26**			4.62**			4.19**		
	Go-by-the-book			Authority					
Age	.03	3.66	.17	.04	5.05*	.20			
Education	.09	8.93**	-.27	.07	4.45*	.20			
Organization dependence on section	.10	1.53	.11	.07	<1	.05			
Technology									
Long-linked	.11	1.36	-.12	.15	10.39**	-.32			
Intensive	.11	<1	.11	.15	<1	.02			
Union density	.21	15.10**	.36	.16	2.10	.14			
Management pressure	.22	<1	-.08	.17	<1	.09			
R^2 adjusted	.17			.12					
F value for the equation	4.78**			3.37**					

Note. $N = 145$. Cum = cumulative; B = standardized regression weight. F tests are for the significance of increments in R^2 .

* $p < .05$. ** $p < .01$.

set of predictors accounted for 18% of the variance. Technology was the only important contributor to supervisor networking, although union density added a small amount of variance. The R^2 for this equation was .19. In addition to technology, age and education were also related to authority, which increased with increments in both. The R^2 was .17.

Using Duncan's multiple-range test, paired comparisons were made between mean leader behavior and authority scores for long-linked, mediating, and intensive technologies. As expected, mean scores for going-by-the-book and emphasizing performance did not differ by technology. Work-group maintenance scores were significantly higher ($p < .001$) in the long-linked and mediating technology units (there were no significant differences between the long-linked and the mediating technology). Network activity scores were significantly higher ($p < .001$) in both the intensive and mediating technologies (no significant differences between these) than in the long-linked technology. Overall authority scores were also significantly higher ($p < .001$) in intensive and mediating technologies (no significant differences between these) than in the long-linked technology.

In addition, paired comparisons were made between work-group maintenance, networking behavior, and leader authority for each of the 10 sections in the long-linked and intensive technologies. For work-group maintenance, mean scores for each one of the intensive technology sections were lower than the scores for each of the long-linked technology sections. Significant differences ($p < .05$) were found between four of the six intensive technology sections and three of the four long-linked technology sections. There were no significant intratechnology-section differences.

For network or linking-pin behavior, supervisors in the two mass-production and the bottling and the packaging sections (long-linked technology) had significantly lower scores ($p < .05$) than did supervisors in the two repair and maintenance units and the one industrial engineering section (intensive technology). Scores for the two research laboratory sections and the print shop were all higher than the scores in the mass-production and packaging sections, but not significantly so. There were no significant differences between section scores within the intensive technology. In the long-linked technology, the plant-cleaning unit was an anomaly. Its supervisors reported significantly more time spent on networking activities than did supervisors in the heavy production and packaging sections.

With respect to leader authority, supervisors in all of the intensive technology sections but one (the print shop) had higher scores than did supervisors in each of the long-linked technology sections. Significant differences ($p < .05$) were found between each of the mass-production and packaging sections and the two repair and maintenance and the industrial engineering sections. There were no significant differences in leader authority scores among the long-linked technology sections. Within the intensive technology sections, the print-shop supervisors had significantly lower ($p < .05$) authority scores than did the repair and maintenance supervisors.

Hypothesis 4 stated that organizational dependence on a unit would require supervisors to push subordinates to perform. This was not supported in the study; quite to the contrary, organizational dependence was negatively related to performance emphasis.

Organizational dependence was not related to the amount of supervisory authority, so there was no support for Hypothesis 5.

Hypothesis 6 stated that union strength would be positively related to supervisors' going-by-the-book behavior. This was supported; union strength contributed substantially to the supervisor consulting and following formal regulations in the treatment of workers. Demographic characteristics also helped determine going-by-the-book; older supervisors paid more attention and younger supervisors paid less attention to rules and regulations. Supervisors with more education were less attentive to going-by-the-book in the treatment of their subordinates. The R^2 for the equation was .22. Union density was not related to the supervisors' overall authority to direct work.

Hypothesis 7, that a management philosophy favoring close and strict supervision would be positively related to such behavior from first-line supervisors, was supported. Management pressure contributed significantly to the performance emphasis dimension beyond the contributions of age (positively) and organizational dependence and technology (negatively). The R^2 for the equation was .17.

Discussion

The purpose of this study was to show how organizational characteristics can act as demands on first-line supervisors to determine their behavior and constrain their authority as leaders. The results clearly support the proposition that the technology in the workplace, labor union strength, and upper management pressure for close and strict supervision relate to the time supervisors spend responding to demands from various constituents. In essence, these factors (and other environmental variables not considered here) define part of the supervisor's job.

In the design of the study, we incurred some costs from the liabilities of creating new measures. The study was independent variable driven, in the sense that we first selected the organizational level variables to be included and then identified the kinds of leader behavior that theoretically should follow from the presence of these organizational constraints. There was no set of already developed scales with known psychometric properties that matched the resulting leader-behavior dimensions. Of the two scales used in this study, the one designed specifically to measure responsive leadership produced measures with satisfactory reliability estimates (work-group maintenance and network development), whereas the other scale (designed originally for use with subordinates) gave less reliable measures. One of the latter measures (performance emphasis), although it was reasonably reliable, turned out to be a deficient criterion because it contained only items of a punitive nature. It was not the most suitable measure with which to examine the effects of organizational dependence on the work unit on leader behavior.

It would have been preferable to use already developed and tested leader-behavior measures. However, most leadership scales are designed for use by subordinates. They are not the same measures when they are transformed for use by supervisors. Scales intended specifically for supervisors, such as the Leader Opinion Questionnaire (Fleishman, 1969), do not measure responsive leadership.

However, with the exception of the performance emphasis scale, we believe there is sufficient evidence from the present

data, in terms of both reliability estimates and the pattern of results from hypothesis testing, to have confidence in the specific measures developed for the study. That does not mean that we consider these measures the ultimate operational definitions of responsive leadership. In retrospect, it would have been helpful to have had criterion data from additional sources, such as ratings of supervisory behaviors by subordinates and superiors, and descriptions of behavior by independent observers. A convergence of multiple measures would have been stronger evidence of validity in this case.

The amount of variance captured by the prediction equation was not large (ranging from 17 to 22 for the different leader-behavior dimensions and authority) but of significance to an argument about the importance of organizational constraints. Also, the results are free from joint method variance because the independent and dependent variable measures came from different sources.

Technology was by far the strongest predictor of responsive leadership and leader authority, suggesting again the argument of a technological imperative in organizations. We argue that its effects on the supervisor are indirect, in the sense that it results in having subordinates with specific characteristics, needs, and demands, and in that it shapes the nature of work that both subordinates and supervisors perform. In this particular study of a production plant, it was also closely linked to the organization's dependence on subunits, which made it difficult to examine dependence as an independent factor. In fact, the use of a single organization in a study of the effects of organizational characteristics constrained the data set somewhat by the high intercorrelations of the situational variables. But a choice of technology in the design of organizations will depend on the organization's mission (if it is a rational choice), so centrality to the core and technology will be related in populations of firms beyond manufacturing.

Union strength showed the expected relation with leader behavior, increasing the attention paid to the union contract and to company rules and regulations in the treatment of workers. It did not act as a curb on supervisory authority, however, beyond the constraints provided by the long-linked technology.

Pressure from upper management for close and strict supervision was significantly related to a rather punitive form of pushing subordinates to work harder—breathing down their necks, chewing them out for poor performance, and so forth. It is not surprising that managerial pressures translate into supervisory behavior, but it is hardly a positive outcome inasmuch as punitive pushing in the long run can lead to an alienated work force.

The issue of responsive versus discretionary leadership requires some attention. We experienced difficulties at the start of this study with the definition of responsive leadership and, in fact, chose as one of our responsive dimensions a behavior that might just as well have been a discretionary one (performance emphasis). It could be argued that almost all leader behavior is a response to organizational demands. For example, facilitating subordinates' work performance through teaching, supplying them with sufficient tools and material to get the work done, or removing obstacles could be a requirement imposed on a leader by organizational characteristics, just as logically as is networking activities or work-group maintenance. We believe that the

distinction between responsive and discretionary leadership, that is, the amount of choice the supervisor has in doing something (see Stewart, 1982), is a useful conceptual tool in leadership research even if it may not be possible to arrive at common operational definitions of the two behavior categories useful in all situations.

Responsive leadership may not be as important as discretionary leadership in influencing subordinates' motivation to work and work attitudes (cf. Hunt & Osborn, 1982), but it certainly is an important part of leadership effectiveness. After all, the role of a work-group leader contains more than supervising subordinates. We believe this study has shown the importance of organizational characteristics as correlates of behavior in, and authority of, the work-group leader role, and we argue for more theoretical and empirical attention to responsive leadership.

References

- Filley, A. C., House, R. J., & Kerr, S. (1976). *Managerial process and organizational behavior*. Glenview, IL: Scott, Foresman.
- Fleishman, E. A. (1969). *Manual for Leadership Opinion Questionnaire* (rev. ed.). Chicago: Science Research Associates.
- Fleishman, E. A., Harris, E. F., & Burt, H. F. (1955). *Leadership and supervision in industry*. Columbus: Ohio State University, Bureau of Educational Research.
- Hall, R. H. (1982). *Organizations: Structure and process* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Hammer, T. H. (1974). The influence of supervisory behavior on subordinates' motivation: A model and a measure of leadership (Doctoral dissertation, University of Maryland, 1973). *Dissertation Abstracts International*, 35, 02-B, 1099.
- Hammer, T. H. (1979). Affirmative action programs: Have we forgotten the first-line supervisor? *Personnel Journal*, 58, 384-389.
- Hammer, T. H., & Dachler, H. P. (1975). A test of some assumptions underlying the path-goal model of supervision: Some suggested conceptual modifications. *Organizational Behavior and Human Performance*, 14, 60-75.
- Homans, G. C. (1965). Effort, supervision and productivity. In R. Dubin, G. C. Homans, F. C. Mann, & D. C. Miller (Eds.), *Leadership and productivity* (pp. 51-67). San Francisco, CA: Chandler.
- House, R. J., & Mitchell, T. R. (1974). Path-goal theory of leadership. *Journal of Contemporary Business*, 3, 81-97.
- Hunt, J. G., & Osborn, R. N. (1982). Toward a macro-oriented model of leadership: An odyssey. In J. B. Hunt, V. Selcaraan, & C. A. Schriesheim (Eds.), *Leadership: Beyond establishment views*. Carbondale, IL: Southern Illinois University Press.
- Jaques, E. (1976). *A general theory of bureaucracy*. New York: Halsted Press.
- Kerr, S., & Jermier, J. M. (1978). Substitutes for leadership: Their meaning and measurement. *Organizational Behavior and Human Performance*, 22, 375-403.
- Kochan, T. A., & Block, R. N. (1977). An interindustry analysis of bargaining outcomes: Primary evidence from 2-digit industries. *Quarterly Journal of Economics*, 91, 431-452.
- Lowin, A., & Craig, J. R. (1968). The influence of level of performance on managerial style: An experimental object-lesson in the ambiguities of correlation data. *Organizational Behavior and Human Performance*, 3, 440-458.
- Pfeffer, J., & Salancik, G. R. (1975). Determinants of supervisory behavior: A role set analysis. *Human Relations*, 28, 1139-1153.
- Pfeffer, J., & Salancik, G. R. (1978). *The external control of organizations: A resource dependence perspective*. New York: Harper & Row.

Psychological Functioning Following an Acute Disaster

Julian Barling

Queen's University, Kingston, Ontario, Canada

Stephen D. Bluen and Rolene Fain

University of the Witwatersrand, Johannesburg, South Africa

We investigated personal and organizational functioning following an acute disaster in an explosives factory in which 14 people were killed and 14 others were injured. Multivariate analyses of covariance (controlling for age and organizational tenure) assessed whether there were any differences between the experimental group (40 individuals physically exposed to the explosion) and two control groups (one from the same site performing a different job, the second from a separate site performing the same job; $n = 76$ and $n = 40$, respectively). During the 2nd week following the blast and 2 months afterward, there were no between-group differences in terms of job satisfaction, organizational commitment, marital satisfaction, or psychological distress. Failure to find any differences was attributed to the acute (as opposed to chronic) nature of the disaster. At both time periods, family support was correlated with personal functioning, whereas supervisory support was associated with job satisfaction; this is discussed in terms of the source of the stressor being consistent with the source of the support and the nature of the outcome.

Research on work stress and its consequences is increasing tremendously (Staw, 1984). Because ethical considerations and problems of generalizability largely rule out the induction of stress in laboratory situations, research on work stress has relied mainly on correlational designs in community (e.g., Motowidlo, Packard, & Manning, 1986) or "at-risk" samples (e.g., Barling & Rosenbaum, 1986). This approach has produced retrospective (e.g., Loo, 1986), cross-sectional (e.g., Motowidlo et al., 1986), and prospective (e.g., Barling & Milligan, 1987) analyses of the correlates of work stress. A problem inherent in most of this research, however, is that correlational or "pre-experimental" designs preclude any true causal inferences regarding the consequences of work stress (cf. Cook & Campbell, 1979).

One way of reducing the limitations of correlational research and the ethical, and content and external validity, issues inherent in analogue research is to study the consequences of work stress in naturalistic field settings that permit the inclusion of relevant control groups. Perhaps the most widely studied work stressor in a naturalistic situation is the disaster at Three Mile Island (TMI; e.g., Chisholm & Kasl, 1982; Chisholm, Kasl, & Eskenazi, 1983; Chisholm, Kasl, & Mueller, 1986; Kasl, Chisholm, & Eskenazi, 1981a, 1981b). These analyses showed that TMI workers experienced negative work-related effects in comparison with control groups in which workers at nuclear power stations did not experience a disaster. Workers at TMI evidenced increased demoralization, job dissatisfaction, and role strains, and negative work-related outcomes were still evident

30 months after the explosion. Chisholm et al. (1986) also showed that social support buffered some negative effects of the TMI disaster.

Findings from studies on the effects of the TMI disaster may not be generalizable to other disasters. Given that disasters vary on a number of dimensions (Baum, Fleming, & Davidson, 1983; Berren, Beigel, & Ghertner, 1986), the generalizability of the TMI phenomenon may be limited to disasters with similar characteristics. The incident at TMI was a man-made disaster, was difficult to predict and of sudden onset, involved a loss of control rather than a lack of control, initially was of unspecified severity, and invoked no visible damage. Also, the disaster at TMI exhibited no "low point" when the effects of the disaster could clearly be said to be finished (Baum et al., 1983; Berren et al., 1986), and as a result, some of the stress and uncertainty associated with the disaster still exists.

One interpretive problem with naturalistic field studies results from the absence of true experimental manipulation, preventing any possibility of sequentially dismantling the various components of the disaster to isolate its true causal components. Therefore, it is not clear which of the characteristics of the TMI disaster led to the negative outcomes reported. Because experimental manipulation of disaster characteristics is not possible, one alternative is to sequentially study several disasters differing in salient characteristics. By comparing two disasters in which only one characteristic varies, the influence of that characteristic could be ascertained.

In the present study, this became possible following a workplace disaster differing from that at TMI. In this situation, three explosions in the nitroglycerine issuing house at the world's largest explosives factory resulted in 14 workers losing their lives and 14 workers suffering injuries of varying severity. Like TMI, this was a man-made disaster of sudden onset that was difficult to predict and that involved a loss of control over the production process. The major difference between the two disasters is that at the explosives plant there was tremendous physical damage and considerable loss of life, and an immediate low point

Portions of this research were supported by grants from the Social Sciences and Humanities Research Council of Canada (Grant 410-85-1139) and Imperial Oil to the first author.

The authors express their appreciation to Karyl MacEwen, Clive Fullagar, Frank Fincham, Sally Grant, and Laurie Pratt for constructive comments at various stages of the research.

Correspondence concerning this article should be addressed to Julian Barling, Department of Psychology, Queen's University, Kingston, Ontario K7L 3N6, Canada.

was reached. As such, we suggest that the incident at the explosives factory was as an acute disaster, whereas the TMI incident can be characterized as a chronic disaster (Pratt & Barling, 1987b). Thus, any differences in outcome between the TMI disaster and the disaster investigated in the present study may be attributable to this acute-chronic distinction, if other explanations (e.g., differences between South African and other national samples) can be excluded. Because of the acute nature of this disaster, and the deaths and physical injuries that resulted, the first hypothesis of the present study is that survivors would experience negative psychological effects.

If this fatal blast is indeed associated with negative psychological outcomes, it is appropriate to investigate variables that moderate its negative effects. Consistent with previous research and theorizing, two general types of coping mechanisms (*viz.*, personality hardiness, and support from friends or supervisor) were hypothesized to moderate the effects of the acute disaster. First, personality hardiness consistently moderates the influence of work stress on health (e.g., Kobasa, 1982; Maddi & Kobasa, 1984) and marital satisfaction (Barling, 1986), at least amongst all-male samples (see MacEwen & Barling, *in press*; Pratt & Barling, 1987a). Personality hardiness includes three personality dimensions: commitment, as opposed to alienation; perceived control, rather than powerlessness; and the perception of events as a challenge rather than a threat. Second, social support moderates the negative influence of work stress on personal and organizational functioning (House, 1981). The effects of social support may depend on its source. Due to suggestions that the source of the support should be consistent with the nature of the stressor (e.g., Beehr, 1985; Kobasa & Puccetti, 1983; Russell, Altmaier, & Van Velzen, 1987), we investigated support from one's supervisor and family.

The present study permits an examination of several additional hypotheses regarding the main or moderating effects of hardiness and support in an acute disaster situation. The second major hypothesis of this study is that personality hardiness is more likely than social support to moderate the influence of stress immediately following the disaster. This is consistent with the suggestion that personality resources will be more effective in acute situations that are of a sudden onset, because they are within the repertoire of the individual and thus readily accessible. On the other hand, it takes time to seek and receive support, and so social support cannot immediately moderate the effects of stress (Hobfall & London, 1986).

It is becoming clear that social support does not always exert positive effects (Beehr, 1985; Kaufmann & Beehr, 1986). Like other findings (e.g., Kaufmann & Beehr, 1986; MacEwen & Barling, *in press*), Hobfall and London (1986) found that support exacerbated the effects of a stressor, as individuals experiencing stress and receiving high levels of support fared significantly worse than did their counterparts who received low levels of support. It has been suggested that positive buffering effects are likely when the source of the stressor and the support are congruent (Beehr, 1985; Russell et al., 1987). Thus, the third major hypothesis in this study is that supervisor support should be more effective because of the work-related nature of the stressor in the present study.

It has also been hypothesized that for social support to be effective, the source of the support must be consistent with the

nature of the outcome (Ganster, Fusilier, & Mayes, 1986; Pratt & Barling, 1987b). We tested this fourth hypothesis by investigating a variety of diverse outcomes. Work-related outcomes included job satisfaction, which is the result of a comparison between the individual's expectations and the current situation (Locke, 1983), and is often an outcome of job stress (e.g., Ganster et al., 1986; Kaufmann & Beehr, 1986). Organizational commitment, which represents the individuals' attachment to the company and is based on an exchange relation between satisfaction of salient needs of the individual and the organization (Mowday, Porter & Steers, 1982), was also measured. We operationally defined personal functioning with two variables, psychological distress and marital satisfaction. Psychological distress reflects the individual's mental health in the occupational setting (Banks et al., 1980) and is associated with the experience of unemployment (Jackson, Stafford, Banks, & Warr, 1983). Both psychological distress and marital dissatisfaction are associated with involvement in a strike situation (Barling & Milligan, 1987), and negative work experiences are associated with marital dissatisfaction (Barling & Rosenbaum, 1986). It was predicted that supervisor support would directly influence organizational commitment and job satisfaction, and would also moderate the effects of the disaster on these two variables. On the other hand, family support would exert a main effect on psychological distress and would moderate the effects of the blast on psychological distress. Personality hardiness would directly affect personal and organizational outcomes equally because it is a stable personality disposition always available to the individual, and hardy individuals exposed to the blast would fare significantly better in terms of psychological functioning than would their nonhardy counterparts.

A fifth and final hypothesis that we considered concerns the duration of any outcome of the disaster. No low point has yet been reached at TMI because community residents remain concerned whether long-term genetic or chromosomal effects remain possible, so it is not surprising that chronic strain is evident (Davidson & Baum, 1986). The disaster we studied, however, was consistent with an acute stressor, where a low point was reached immediately (Pratt & Barling, 1987b). Loo (1986) indicated that the consequences of acute stressors are experienced immediately and dissipate after a few days. Therefore, it was hypothesized that any negative effects would be more likely to emerge immediately following the blast and to dissipate within 2 months thereafter.

Method

Subjects and Setting

On February 14, 1985, an explosion occurred in the nitroglycerine issuing house of the world's largest explosives factory just outside Johannesburg, South Africa. Even though all work in this issuing house was robotically controlled, the initial blast involving 1.5 tons of explosives in this unit immediately set off two sympathetic blasts in adjacent units, one involving an ammonium nitrate store. The effects of the blast were devastating. As an indication thereof, note that despite the fact that the buildings housing the various units were geographically separated and designed so that the effects of an explosion would be deflected upwards, the units in which the blasts occurred were physically annihilated, windows were shattered for miles around, shock waves from the blast were

felt in a radius of 15–20 miles, and a large pall of black smoke could be seen rising from the scene after the blast. In all, 14 workmen were killed and 14 sustained minor or major injuries, mainly involving burns. Of those killed, 3 died instantly, 6 died in the first few days following the blast, and the bodies of the remaining 5 were never recovered.

The experimental group included the 40 survivors employed in the blasting department who were present at the scene of the blast. Two control groups were formed in this study. The first (Control Group 1) included the 76 White employees who were working in an ammonia plant in the same organization and at the same site at the time of the blast. They all heard the blast and felt the shock waves of the blast, but were in no physical danger themselves. However, their proximity to the blast itself may increase their similarity to the experimental group, and therefore, reduce their value as a control. As a result, a second control group (Control Group 2) was used, which included 40 individuals in a nitroglycerine issuing department at a different site, owned and operated by the same company but where no explosion had occurred.

Initially, there were 4 Black employees in the experimental group and 3 in the control groups. All 7 were excluded from the study, however, as there is some evidence that the work-related needs and attitudes of Blacks and Whites in South Africa differ (Fullagar & Barling, 1986), and the small number of Blacks in these groups would not have allowed a statistical comparison of race as a moderator of the effects of the blast.

The three groups then differed in terms of age (experimental group, *M* age = 37 years; Control Group 1, *M* age = 34 years; Control Group 2, *M* age = 41 years), and length of tenure with the organization (experimental group, *M* years = 11; Control Group 1, *M* = 9 years; Control Group 2, *M* = 14 years). However, there were no differences between the three groups in terms of monthly salary (*M* = R2118; i.e., ±\$900) or educational level (43% had completed grade 12; 57% had completed some tertiary education). Because supervisory status may influence an individual's response to a work-related disaster (Chisholm & Kasl, 1982), only employees at the nonsupervisory level were included in this study.

Of the initial sample, 63% agreed to participate at the 2-month follow-up phase. There were no differences between those who agreed to respond and those who did not, in terms of demographic characteristics, moderator, or dependent variables immediately following the blast (*p* > .05).

Procedure

During the 2nd week following the disaster, and again 2 months after the disaster, all of the subjects completed a questionnaire package. Subjects were assured of the confidentiality of their responses.

Measuring Instruments

All of the data were collected via self-reports obtained from questionnaires. Questionnaires were selected according to whether they had proven to be psychometrically acceptable. Descriptive statistics, internal and temporal consistency, and intercorrelations of all variables included are presented in Table 1.

Dependent variables. Two different aspects of psychological functioning were investigated, namely, work and personal functioning. To assess job satisfaction, Warr, Cook, and Wall's (1979) 15-item job satisfaction measure, which was developed specifically to be appropriate for blue-collar workers, was administered. However, a 3-point rating scale (*unhappy, not sure, happy*) was used instead of the 7-point rating scale, to facilitate understanding. Mowday, Porter and Steers's (1982) 9-item short form of the Organizational Commitment Questionnaire was used to assess respondents' attachment to the company. Again, a 3-point rating scale (*disagree, not sure, agree*) was used. The 12-item General Health Questionnaire (GHQ; Goldberg, 1972) assessed psychological

Table 1
Descriptive Data and Intercorrelations of Variables

Variable	<i>M</i>	<i>SD</i>	α	Test-retest	1	2	3	4	5	6	7	8	9
1. Tenure	12.06	9.04	—	—	—	.80	.18	.01	.07	.31**	.09	.07	-.06
2. Age	37.89	10.44	—	—	.80	—	.25**	.10	-.01	.30**	.18*	.03	-.03
3. Supervisor support	18.89	4.35	.62	.67	.02	.10	—	.13	.19*	.33**	.50**	-.25**	.13
4. Family support	50.26	5.96	.82	.80	-.02	.07	.04	—	.33**	.34**	.22*	-.27*	.70**
5. Personality hardness	30.69	5.74	.71	.73	-.07	-.01	.34**	.25**	—	.28**	.32**	-.39*	.45**
6. Organizational commitment	20.94	4.84	.87	.84	.21**	.25**	.22*	.26**	.24**	—	.52**	-.20*	.27**
7. Job satisfaction	36.73	6.04	.81	.69	.11	.18*	.51**	.20**	.40**	.66**	—	-.24**	.24*
8. Psychological distress	10.65	4.27	.76	.66	.15*	.12	-.02	-.32**	-.25**	-.28**	-.35**	—	-.32**
9. Marital satisfaction	113.46	35.04	—	.54	-.05	.01	.13	.65**	.33**	.25**	.27**	-.35**	—

Note. Data during the 2nd week following the blast are presented below the diagonal (*n* = 156, except for marital satisfaction, where *n* = 102 because of nonmarried individuals); data 2 months following the blast are presented above the diagonal (*n* = 122, except for marital satisfaction, where *n* = 73).
p* < .05. *p* < .01.

distress. This scale provides useful data within the occupational context (Banks et al., 1980) and is scored such that a high score reflects psychological distress. Locke and Wallace's (1959) 15-item Short Marital Adjustment Test (SMAT) was used to assess marital satisfaction. The SMAT remains the most frequently used scale of marital satisfaction (O'Leary & Turkewitz, 1978) and reliably discriminates between clinically distressed and satisfactory marital relationships (O'Leary & Arias, 1987).

Moderator variables. To assess personality hardiness, Kobasa's (personal communication, 1982) 20-item shortened form of the original hardiness scale was administered. Although no data was provided initially for this scale, subsequent research has shown that it exhibits adequate internal and temporal consistency (Barling, 1986; MacEwen & Barling, in press; Pratt & Barling, 1987a). To assess family support, Procidano and Heller's (1983) 20-item Perceived Social Support From Family scale was also completed by all subjects. Finally, Moos's (1981) 9-item Supervisor Support subscale from the Work Environment Scale-Real Form was completed.

Data Analysis

A test of the first hypothesis—namely, that exposure to an acute disaster will result in negative psychological effects—requires an evaluation of whether significant differences emerge between the experimental group and either or both of the two control groups. The second, third, and fourth hypotheses revolve around the issue of whether support or personality hardiness, or both, moderate any detrimental psychological effects of exposure to the acute disaster. Within a univariate analysis of variance (ANOVA) design, an assessment of any moderating effects can be undertaken through an inspection of the interaction term. In general, for example, family or supervisory support, or both, would serve as a moderator variable if those individuals who were exposed to the blast and who received high family or supervisory support, or both, fared significantly better than their counterparts who were exposed to the disaster but were not in receipt of such support. Likewise, personality hardiness would fulfill a moderating role if hardy individuals exposed to the blast functioned significantly better thereafter than did nonhardy individuals (cf. Thoits, 1982). For the purpose of the ANOVA design, median splits were performed on the moderator variables (supervisor support, $mdn = 20$; family support, $mdn = 51$; personality hardiness, $mdn = 39$).

Results

Because the experimental and control groups differed in length of service in the company and in age, these two variables were controlled statistically in all of the analyses. A series of 3×2 (Group Status \times Moderator Variable) multivariate analyses of covariance (MANCOVAs) were computed to assess whether there were significant differences between the control and experimental groups. Multivariate analyses were computed because of the numerous significant correlations between the dependent variables within the 2nd week following the blast and 2 months thereafter (see Table 1), and because of the large number of analyses computed. In all of these MANCOVAs, the Pillai-Bartlett trace F approximation was used to test for significant between-group differences.

The number of subjects for whom data was available for each of the four dependent variables differed. First, not all of the subjects were married; hence, the SMAT could not be completed by all subjects. Second, there were instances in which incomplete data led to a particular questionnaire being unusable for

specific subjects. Consequently, ANOVAs are based on a somewhat different sample size.

Psychological Functioning During the Second Week After the Blast

There were no significant differences between the control and experimental groups on any of the four dependent variables (viz., job satisfaction, organizational commitment, marital satisfaction, or psychological distress) during the 2nd week following the blast ($p > .05$), after controlling for age and organizational tenure. At the same time, the moderating role of supervisor support, family support, and personality hardiness were investigated. Neither supervisory support, family support, nor personality hardiness emerged as significant moderator variables ($p > .05$) in the multivariate analyses.

At that stage, however, supervisory support, $F(4, 89) = 5.69$, $p < .01$; family support, $F(4, 89) = 9.01$, $p < .01$; and personality hardiness, $F(4, 89) = 4.94$, $p < .01$, all exerted significant multivariate main effects. To assess which of the univariate dependent variables contributed to these multivariate effects, univariate analyses were computed. However, because this involves computation of numerous F tests, the probability of Type I errors increases. Thus, Bonferroni's procedure (Larzelere & Mulaik, 1977) was used to adjust the significance level. This resulted in adopting a significance level of .003 in all analyses (i.e., alpha level/number of F tests).

Supervisory support did not predict psychological distress, $F(1, 123) = 0.10$; marital satisfaction, $F(1, 89) = 0.67$; or organizational commitment, $F(1, 123) = 4.39$, but was associated with job satisfaction, $F(1, 123) = 36.37$, $p < .003$, after controlling for age and tenure in each case. Family support predicted marital satisfaction, $F(1, 78) = 38.37$, $p < .003$, and psychological distress, $F(1, 83) = 8.65$, $p < .003$, but neither organizational commitment, $F(1, 122) = 4.21$, nor job satisfaction, $F(1, 123) = 4.01$. Personality hardiness was not significantly associated with organizational commitment, $F(1, 104) = 7.22$, but was associated with job satisfaction, $F(1, 104) = 16.50$, $p < .003$; marital satisfaction, $F(1, 78) = 14.13$, $p < .003$; and psychological distress, $F(1, 104) = 20.98$, $p < .003$. The adjusted mean scores for all these analyses are presented in Table 2.

Psychological Functioning Two Months Following the Blast

This pattern of results was largely replicated 2 months following the blast. After again controlling for the influence of age and tenure in the organization, MANCOVAs showed that there were no significant differences between the experimental and control groups on any of the four dependent variables. In addition, neither personality hardiness nor family support or supervisory support (measured 2 months following the blast) interacted with group status. However, there was a significant multivariate main effect for supervisory support, $F(4, 59) = 7.44$, $p < .001$; family support, $F(4, 54) = 6.40$, $p < .001$; and personality hardiness, $F(4, 55) = 4.07$, $p < .005$.

Univariate F tests were again computed, and Bonferroni's procedure again resulted in a significance level of .003 being accepted to reduce the probability of Type I errors because of

Table 2

Dependent variable	Group status			Supervisor support		Family support		Personality hardness	
	Control 1	Control 2	Experimental	High	Low	High	Low	High	Low
Organizational commitment	20.56	21.08	20.97	21.78	19.97	22.20	19.55	22.12	19.72*
Job satisfaction	35.25	37.45	37.18	39.85	34.02*	37.75	36.27	38.72	34.13
Psychological distress	10.69	10.50	10.26	10.58	10.33	9.70	12.56*	9.11	12.99*
Marital satisfaction	101.30	115.63	120.47	114.01	110.55	129.73	91.76	120.07	102.98

Note. $N = 156$.

² Controlling for age and length of service in the organization.

* Indicates a significant difference, with $p < .003$.

the numerous F tests computed. The main effects for supervisory support on psychological distress, $F(1, 88) = 3.89$, and marital satisfaction, $F(1, 68) = 0.10$, were not significant. Job satisfaction, however, was associated with supervisory support, $F(1, 88) = 26.20, p < .001$, as was organizational commitment, $F(1, 88) = 8.68$. Family support was significantly associated with marital satisfaction, $F(1, 63) = 16.28, p < .001$, and psychological distress, $F(1, 83) = 8.19, p < .003$, but not with job satisfaction, $F(1, 81) = 2.46$, or organizational commitment, $F(1, 82) = 4.55$. Personality hardness predicted job satisfaction, $F(1, 83) = 8.2, p < .003$, and organizational commitment, $F(1, 83) = 10.78, p < .002$, but neither marital satisfaction, $F(1, 64) = 3.32$, nor psychological distress, $F(1, 83) = 6.05$. These main effects were all in the predicted direction, and the adjusted mean scores are presented in Table 3.

Discussion

The major finding of this research is that workers directly exposed to the blast evidenced no detrimental psychological effects within the 2nd week following the blast or 2 months thereafter, as compared with the two control groups. Thus, no support emerged either for the first or the fifth hypothesis. In attempting to understand the reasons for this seemingly counterintuitive finding, note that some degree of confidence can be placed on the results from this naturalistic quasi-experimental study. Even though the presence of pretest data must obviously be sacrificed in such disaster studies, the experimental group did not differ from either of two control groups across any of the four dependent variables at either of the two testing phases. Thus, the results of the present study do not replicate the analyses conducted on work-related or personal consequences of the TMI disaster (Chisholm & Kasl, 1982; Chisholm et al., 1983, 1986).

We suggest that the major difference between the nature of the stressor in these two disasters might account for the fact that no detrimental psychological effects emerged following the fatal blast. Specifically, unlike the disaster at TMI that continues to exert detrimental effects because of its chronic nature (Chisholm et al., 1983, 1986; Davidson & Baum, 1986), the fatal blast investigated in this research can be characterized as an acute stressor (Pratt & Barling, 1987b): It was of sudden onset, high intensity, extremely low frequency, and a low point was reached immediately after the fatal blast. The fact that no negative psychological functioning was apparent in the experimental group in the present study is consistent with other findings suggesting that most of the detrimental effects of acute work stressors dissipate within 3 days of the stressor (Loo, 1986). Consistent with Loo's findings, it would be premature to suggest that acute work-related stressors exert no negative effects because the first measurement in the present study occurred during the 2nd week following the fatal blast. Rather, it remains possible that negative symptoms were experienced immediately following the blast but were no longer apparent a few days thereafter. Future research should investigate this possibility, which could not be assessed in this study due to logistic difficulties.

Two alternative explanations should be considered. First, it has been suggested that negative functioning following a disaster is more likely to become apparent if psychiatric interviews are

Table 3
Adjusted Mean Scores and Variance Accounted for Two Months Following the Blast^a

Dependent variable	Group status			Supervisor support			Family support			Personality hardness		
	Control 1	Control 2	Experimental	Variance (%)		High	Variance (%)		High	Variance (%)		Low
Organizational commitment	20.43	21.11	21.42	11.0	22.32	19.42	17.5	21.94	22.66	19.24*	19.0	
Job satisfaction	36.56	36.85	37.70	5.4	39.89	34.23*	26.7*	37.98	39.14	35.17*	13.1	
Psychological distress	10.22	10.80	9.40	2.7	9.28	11.12	5.3	8.90	8.95	11.17	7.6	
Marital satisfaction	119.23	102.37	108.68	5.9	111.68	114.25	1.4	122.45	116.62	104.07*	5.3	

Note. *N* = 122.
^a Controlling for age and length of service in the organization.
* Indicates a significant difference with *p* < .003.

used to gather information rather than standardized psychological tests, the assumption being that the former are more sensitive to the nature of psychiatric symptoms (Perry, 1979). However, this argument is not supported. (a) Using the data obtained at TMI, it can be seen that consistent negative effects emerged for 58 months following the disaster using standardized questionnaires (Davidson & Baum, 1986). (b) Dollinger, O'Donnell, and Staley (1984) have shown that the effects of an acute stressor can be isolated whether interview or survey methodologies are used.

Second, the argument that standardized psychological questionnaires may fail to detect psychiatric symptoms following a disaster can also be refuted because of the nature of two of the four questionnaires used. Specifically, the SMAT and the GHQ are particularly sensitive measures that can detect clinical symptoms. Locke and Wallace's (1959) SMAT remains the most widely used measure of global marital satisfaction in the clinical field (O'Leary & Turkewitz, 1978), and scores below 100 consistently denote a marriage that is clinically at risk (cf. Barling & Rosenbaum, 1986). Yet individuals who were physically exposed to the blast yielded marital satisfaction scores that clearly place them in the range of maritally satisfied couples, both 2 weeks and 2 months following the disaster (*M* = 120.47 and *M* = 108.68, respectively). Likewise, the GHQ is a sensitive measure of occupational mental health (Banks et al., 1980), and research has shown that psychological distress associated with a strike (Barling & Milligan, 1987) or unemployment (e.g., Jackson et al., 1983) can be detected with the GHQ. The GHQ scores of individuals in the experimental group in the 2nd week following the blast or 2 months thereafter (*M* = 10.26 and 9.40, respectively) is within the normal range of other studies (cf. Banks et al., 1980; Jackson et al., 1983) and did not differ from either of the two control groups used in the present study.

It remains possible, however, that even though two measures of each of organizational and personal functioning were obtained, areas of functioning most likely to be affected following a disaster were not considered. Lifton and Olson (1986) suggested that the content of dreams might well change following a disaster, and that there may be significant increases in psychic numbing, death anxiety, and death guilt. The questionnaires used in the present study did not address these possible outcomes. In addition, whereas the present research concentrated on somewhat general aspects of psychological functioning, Dollinger et al's (1984) results suggested that highly specific fears may follow disasters. As a result, it is suggested that, where possible, future research should be conducted immediately following an acute disaster (i.e., within a few days) and should focus on general experiences such as those identified by Lifton and Olson (1986) or on fears that might be considered to be highly specific to the nature of the disaster.

The second major finding of the present study concerns the role of hypothesized moderator variables. A number of issues emerge. First, contrary to the second, third, and fourth hypotheses, no moderating effects emerged. Although it is becoming apparent that social support does not necessarily exert positive buffering effects (e.g., Ganster et al., 1986; Hobfall & London, 1986; Kaufmann & Beehr, 1986; MacEwen & Barling, in press; Pratt & Barling, 1987a), the failure to find evidence for a moderating role in this study is probably due to the fact that at both

testing phases, individuals in the experimental group were not functioning at a level different from those in either of the two control groups. Any evidence for a moderating role for social support assumes the existence of high levels of stress (Thoits, 1982). Certainly, scores on both the GHQ and the SMAT for workers in the experimental group at both testing phases are not significantly different from those experiencing satisfactory marital relationships (see Barling & Rosenbaum, 1986) or psychological well-being (Banks et al., 1980; Jackson et al., 1983). It is argued that the absence of stress related to the disaster for workers in the experimental group also accounts for the failure of personality hardiness to exert a moderating effect. Thus, the second and third hypotheses—namely, that personality hardiness would exert immediate moderating effects, whereas those for social support would be delayed—could not be assessed because of the absence of any stress effects.

A further, although not necessarily alternative, explanation for the findings regarding supervisor and family support concerns the nature of the support measured. Only emotional support from the supervisor and family was measured in this study. Yet recent research results show that although emotional support is usually positively associated with various outcome measures, emotional support either exacerbates (e.g., Hobfall & London, 1986; Kaufmann & Beehr, 1986; Kobasa & Puccetti, 1983; MacEwen & Barling, in press; Pratt & Barling, 1987a) or exerts no moderating effect (Ganster et al., 1986) on diverse measures of strain. Even though most studies have focused on emotional support—presumably because of the existence of appropriate questionnaires—future research should focus on instrumental and informational support (e.g., House, 1981), particularly inasmuch as research at TMI has suggested the moderating role of informational support (Chisholm et al., 1986).

Consistent with the fourth hypothesis, the results of the present research strengthen the suggestion that the source of the support should be congruent with the nature of the outcome (Ganster et al., 1986; Pratt & Barling, 1987b). At both testing periods, family support was associated positively with personal functioning, whereas supervisor support was positively correlated with work-related variables. However, family support predicted personal outcomes, and supervisor support predicted job satisfaction, irrespective of prior levels of stress. In addition, the positive benefits accruing to personality hardiness were not limited to either personal or organizational functioning. Most previous research has focused on the role of support and hardiness during normal situations. The results of the present study suggest that social support and hardiness continue to exert main effects on relevant outcomes during disaster situations.

A major strength of the present study is its inclusion of two control groups that permit an evaluation of plausible explanations concerning the absence of any detrimental effects during the 2nd week following the blast, and 2 months thereafter. A further strength concerns the use of a specific rationale for predicting when the effects of the disaster might emerge. A potential limitation concerns the exclusive reliance on self-report measures. However, given the nature of research in disaster situations, desirable experimental procedures (including the use of pretest measures) may often need to be sacrificed in an effort to test the victims as soon as possible. Even so, it remains questionable whether the use of self-report in this research is a limi-

tation for two reasons, both of which are related to primary concerns about the use of single source data sets. First, previous research on disasters has shown that the results of self-report data (as opposed to behavioral measures) do not overestimate the effects of the disaster (e.g., Davidson & Baum, 1986; Dollinger et al., 1984). Second, to remain congruent with the experience of the stressor, victims might be more likely to represent themselves as psychologically impaired. This possibility can be excluded in this study as individuals in the experimental group did not function any worse than those in either of the two control groups on the four outcome variables measured.

It remains for future researchers investigating the acute effects of work-related disasters to focus on other outcome measures (e.g., psychic numbing, grief, nightmares, and fears specific to the disaster) at time periods different from those used in the present study. Where possible, data should be gathered within the first 3 days following the disaster and perhaps up to a year or more following the disaster. This would enable an assessment of whether there are immediate negative effects that dissipate within a few days. Alternatively, perhaps no negative effects emerged in this study because it takes longer than 2 months before any negative outcomes become apparent.

References

- Banks, M. J., Clegg, C. W., Jackson, P. R., Kemp, N. J., Stafford, E. M., & Wall, T. D. (1980). The use of the General Health Questionnaire as an indicator of mental health in occupational settings. *Journal of Occupational Psychology*, 53, 187–194.
- Barling, J. (1986). Interrole conflict and marital functioning amongst employed fathers. *Journal of Occupational Behaviour*, 7, 1–8.
- Barling, J., & Milligan, J. (1987). Some psychological consequences of striking: A six-month longitudinal study. *Journal of Occupational Behaviour*, 8, 127–137.
- Barling, J., & Rosenbaum, A. (1986). Work stressors and wife abuse. *Journal of Applied Psychology*, 71, 384–386.
- Baum, A., Fleming, R., & Davidson, L. M. (1983). Natural disasters and technological catastrophe. *Environment and Behavior*, 15, 333–354.
- Beehr, T. A. (1985). The role of social support in coping with organizational stress. In T. A. Beehr & R. S. Bhagat (Eds.), *Human stress and cognition in organizations: An integrated perspective* (pp. 375–400). New York: Wiley.
- Berren, M. R., Beigel, A., & Ghertner, S. (1986). A typology for the classification of disasters. In R. H. Moos (Ed.), *Coping with life crisis: An integrated approach* (pp. 295–305). New York: Plenum Press.
- Chisholm, R. F., & Kasl, S. V. (1982). The effects of work site, supervisory status and job function on nuclear workers' responses to the TMI accident. *Journal of Occupational Behaviour*, 3, 39–62.
- Chisholm, R. F., Kasl, S. V., & Eskenazi, L. (1983). The nature and predictors of job related tension in a crisis situation: Reactions of nuclear workers to the Three Mile Accident. *Academy of Management Journal*, 26, 385–405.
- Chisholm, R. F., Kasl, S. V., & Mueller, L. (1986). The effects of social support on nuclear worker responses to the Three Mile Accident. *Journal of Occupational Behaviour*, 7, 179–194.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.
- Davidson, L. M., & Baum, A. (1986). Chronic stress and post-traumatic stress disorders. *Journal of Consulting and Clinical Psychology*, 54, 303–308.
- Dollinger, S. J., O'Donnell, J. P., & Staley, A. A. (1984). Lightning-strike

- disaster: Effects on children's fears and worries. *Journal of Consulting and Clinical Psychology*, 52, 1028-1038.
- Fullagar, C., & Barling, J. (1986). *A longitudinal test of a model of the antecedents and consequences of union commitment*. Manuscript submitted for publication, Queen's University, Kingston, Ontario, Canada.
- Ganster, D. C., Fusilier, M. R., & Mayes, B. T. (1986). Role of social support in the experience of stress at work. *Journal of Applied Psychology*, 71, 102-110.
- Goldberg, D. P. (1972). *The detection of psychiatric illness by questionnaire*. London: Academic Press.
- Hobfall, S. E., & London, P. E. (1986). The relationship of self-concept and social support to emotional distress among women during war. *Journal of Social and Clinical Psychology*, 2, 189-203.
- House, J. S. (1981). *Work stress and social support*. Reading, MA: Addison-Wesley.
- Jackson, P. R., Stafford, E. M., Banks, M. J., & Warr, P. B. (1983). Unemployment and psychological distress in young people: The moderating role of employment commitment. *Journal of Applied Psychology*, 68, 525-535.
- Kasl, S. V., Chisholm, R. F., & Eskenazi, B. (1981a). The impact of the accident at the Three Mile Island on the behavior and well-being of nuclear workers: Part 1. Perceptions and evaluations, behavioral responses, and work-related attitudes and feelings. *American Journal of Public Health*, 71, 472-483.
- Kasl, S. V., Chisholm, R. F., & Eskenazi, B. (1981b). The impact of the accident at the Three Mile Island on the behavior and well-being of nuclear workers: Part 2. Job tension, psychophysiological symptoms, and indices of distress. *American Journal of Public Health*, 71, 484-495.
- Kaufmann, G. M., & Beehr, T. A. (1986). Interactions between job stressors and social support: Some counterintuitive results. *Journal of Applied Psychology*, 71, 522-526.
- Kobasa, S. C. (1982). The hardy personality: Toward a social psychology of stress and health. In G. S. Suls & J. A. Sanders (Eds.), *Social psychology of health and illness* (pp. 3-32). Hillsdale, NJ: Erlbaum.
- Kobasa, S. C., & Puccetti, M. C. (1983). Personality and social resources in stress resistance. *Journal of Personality and Social Psychology*, 45, 839-850.
- Larzelere, R. E., & Mulaik, S. A. (1977). Single sample tests for many correlations. *Psychological Bulletin*, 84, 557-569.
- Lifton, R. J., & Olson, E. (1986). The human meaning of total disaster: The Buffalo Creek disaster. In R. H. Moos (Ed.), *Coping with life crisis: An integrated approach* (pp. 307-321). New York: Plenum Press.
- Locke, E. A. (1983). The nature and causes of job satisfaction. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 1297-1350). New York: Wiley.
- Locke, H. J., & Wallace, K. M. (1959). Short marital adjustment and prediction tests: Their reliability and validity. *Marriage and Family Living*, 21, 251-255.
- Loo, R. (1986). Post-shooting stress reactions among police officers. *Journal of Human Stress*, 12, 27-31.
- MacEwen, K. E., & Barling, J. (in press). Interrole conflict, family support and marital functioning of employed mothers: A short-term longitudinal study. *Journal of Occupational Behaviour*.
- Maddi, S. R., & Kobasa, S. C. (1984). *The hardy executive: Health under stress*. Homewood, IL: Dow Jones-Irwin.
- Moos, R. H. (1981). *Work Environment Scale-Form R*. Palo Alto, CA: Consulting Psychologists Press.
- Motowidlo, S. J., Packard, J. S., & Manning, M. R. (1986). Occupational stress: Its causes and consequences for job performance. *Journal of Applied Psychology*, 71, 618-629.
- Mowday, R. T., Porter, L. T., & Steers, R. M. (1982). *Employee-organization linkages: The psychology of commitment, absenteeism and turnover*. New York: Academic Press.
- O'Leary, K. D., & Arias, I. (1987). Marital assessment in clinical practice. In K. D. O'Leary (Ed.), *Assessment of marital discord: An integration of research and clinical practice* (pp. 287-312). Hillsdale, NJ: Erlbaum.
- O'Leary, K. D., & Turkewitz, H. (1978). Methodological errors in marital and child treatment research. *Journal of Consulting and Clinical Psychology*, 47, 747-758.
- Perry, R. W. (1979). Detecting psychopathological reactions to natural disaster: A methodological note. *Social Behavior and Personality*, 7, 173-177.
- Pratt, L. I., & Barling, J. (1987a). *Development and test of a moderated-mediation model of the relationship between supervisory support, role stressors, and burnout*. Manuscript submitted for publication, Queen's University, Kingston, Ontario, Canada.
- Pratt, L. I., & Barling, J. (1987b). Differentiating between daily events, acute, and chronic stressors: A framework and its implications. In J. R. Hurrell, L. R. Murphy, S. L. Sauter, & C. L. Cooper (Eds.), *Occupational stress: Issues and developments in research* (pp. 41-53). London: Taylor & Francis.
- Procidano, M. E., & Heller, K. (1983). Measures of perceived social support from friends and from family: Three validation studies. *American Journal of Community Psychology*, 11, 1-25.
- Russell, D. W., Altmaier, E., & Van Velzen, D. (1987). Job-related stress, social support, and burnout among classroom teachers. *Journal of Applied Psychology*, 72, 269-274.
- Staw, B. M. (1984). Organizational psychology: A review and reformulation of the field's outcome variables. *Annual Review of Psychology*, 35, 627-666.
- Thoits, P. A. (1982). Conceptual, methodological, and theoretical problems in studying social support as a buffer against life stress. *Journal of Health and Social Behavior*, 23, 145-159.
- Warr, P. B., Cook, J., & Wall, T. D. (1979). Scales for the measurement of work attitudes and aspects of psychological well-being. *Journal of Occupational Psychology*, 52, 129-148.

Received March 13, 1987

Revision received June 4, 1987

Accepted June 8, 1987 ■

Correlation of Eyewitness Accuracy and Confidence: Optimality Hypothesis Revisited

Robert K. Bothwell
Florida State University

Kenneth A. Deffenbacher
University of Nebraska at Omaha

John C. Brigham
Florida State University

Inasmuch as a completely satisfactory estimate of effect size for the eyewitness accuracy-confidence relation does not exist, we conducted a meta-analysis of 35 staged-event studies. Estimated $r = .25$ ($d = .52$), with a 95% confidence interval of .08 to .42. Sampling error accounted for 52% of the variation in r , leaving room for measurement error and possibly moderator variables to account for the remaining variation. Further analysis identified duration of target face exposure as a moderator variable, providing support for Deffenbacher's (1980) optimality hypothesis. When corrected for the attenuating effect of sampling error in the accuracy-confidence correlations, the correlation of exposure duration and the accuracy-confidence correlation was .51: Longer exposures allowed for greater predictability of accuracy from confidence. Even though correction for unreliability in the confidence measure produces a higher estimate of the population correlation of accuracy and confidence, .34, one must be cautious in assessing the utility of confidence for predicting accuracy in actual cases.

Trial attorneys have known for a long time that jurors rely heavily on a witness's degree of confidence as an index of his or her accuracy. In fact, this reliance on confidence to predict accuracy has been formally enshrined in criteria used by the U.S. judiciary for deciding the trustworthiness of eyewitness testimony (*Neil v. Biggers*, 1972). Indeed, there is empirical evidence of a heavy use of the eyewitness confidence criterion. A series of studies by Wells and his colleagues (e.g., Wells, Lindsay, & Ferguson, 1979) showed juror perceptions of witness confidence as accounting for as much as 50% of the variance in juror judgments as to witness accuracy. In a more recent study, Cutler, Penrod, and Stuve (in press) found that of 10 witness and identification variables, only witness confidence had a reliable effect on jurors' judgments as to the probability that the identification had been correct. It is well to ask whether this reliance on the confidence criterion is well placed, however.

Deffenbacher (1980) reviewed a rather heterogeneous collection of studies done over the previous 80 years that had assessed the accuracy-confidence relation in eye- and earwitnesses. He concluded that the strong faith in the adequacy of certainty as a predictor of accuracy was not at all supported by the evidence. Statistical support was found, though, for the optimality hypothesis. The notion here is that predictability of accuracy from overtly expressed confidence varies directly with the optimality of information-processing conditions during encoding of the

witnessed event, memory storage, and testing of the witness's memory.

A problem with Deffenbacher's (1980) review, however, is that no quantitative estimate was provided of the accuracy-confidence correlation across studies reviewed. Indeed, several of the studies included in the review did not report the actual magnitude of the correlation, reporting only that it was "not significant." Wells and Murray (1984) have at least partly mitigated this difficulty in their review of some 31 studies, several of which were published after Deffenbacher's (1980) review. Their estimate of the average r was .07 ($d = .14$). An effect of this size is clearly negligible.

Before accepting Wells and Murray's (1984) estimate of effect size as valid, though, two difficulties with their review should be addressed. For one thing, studies sampled by Wells and Murray were not homogeneous with respect to nature of the task required of the witness, the measure of memory accuracy taken, and the manner in which the accuracy-confidence correlation was computed. Such heterogeneity may well have had the effect of adding unnecessary variation in the size of the correlations sampled. The heterogeneity certainly makes generalization of Wells and Murray's aggregate correlation more hazardous.

Some 22 of the studies reviewed by them measured only recognition memory of once-seen unfamiliar faces. Four studies, on the other hand, assessed recognition memory of rather different stimuli: a critical traffic sign, an unfamiliar voice, a set of celebrity faces, and different views of a very familiar face, the observer's own face. Yet another four studies correlated a recall measure of memory reliability with witness confidence. Recall was assessed not just for characteristics of the "criminal" but also for characteristics of the "crime" setting and actions taken there. Meanwhile, there is accumulating evidence that recall or reconstructive measures of facial memory may be at least par-

Correspondence concerning this article should be addressed to Robert Bothwell, who is now at the Department of Psychology, Pan American University, Edinburg, Texas 78539, or to Kenneth Deffenbacher, Department of Psychology, University of Nebraska at Omaha, Omaha, Nebraska 68182-0274.

tially independent of recognition measures. Jenkins and Davies (1985) and Pigott and Brigham (1985) found no relation between recall and recognition measures of face memory, whereas Wells (1985) found a significant but very modest correlation, .27. It may well be that encoding strategies that facilitate verbal recall do not facilitate facial recognition (Wells & Hryciw, 1984). Finally, one study cited by Wells and Murray (1984), Brigham, Maass, Martinez, and Whittenberger (1983), reported an average intraindividual (within-subjects) correlation of accuracy and confidence, rather than the typically computed interindividual correlation (between subjects).¹

The other problem with the Wells and Murray (1984) estimate of effect size is that it was not a product of the most powerful meta-analysis procedures available. Wells and Murray presented a list of studies, indicating whether each produced a significant correlation of accuracy and confidence. They did not specify how they estimated the overall effect size. Presumably, it was not based on averaging of actual numerical values of r , inasmuch as at least three studies in their sample did not report a numerical value of r or a value of t or F that would allow derivation of a point-biserial correlation coefficient. Perhaps effect size was estimated from the proportion of positive significant results (Hedges & Olkin, 1980).

In any event, a state-of-the-art procedure should be used such as that specified by Hunter, Schmidt, and Jackson (1982). This procedure, followed in the present study, emphasizes averaging of r s weighted by sample size, correction of the variance in r for sampling error, measurement error, and range effects, and, finally, construction of confidence limits for the estimated effect size. With this technique, if sufficient variance in r remains after correction for statistical artifacts, a search for moderator variables is carried out.

Should a search for moderator variables become necessary after the current meta-analysis, we decided that the first candidate for test should be duration of target face exposure (opportunity to observe). This selection was made for two reasons. It has been an important variable to the judicial system for evaluating the trustworthiness of eyewitness testimony (e.g., *Neil v. Biggers*, 1972), and it has already shown promise as a variable moderating the accuracy–confidence relation. Shapiro and Penrod (1984) found an r of .48 between target exposure duration and the accuracy–confidence correlation.

At this juncture, however, their result should be treated as preliminary in that the main purpose of their study was to determine average effect size for some 19 variables, effect size on both face recognition hits and false alarm responses. Shapiro and Penrod's (1984) meta-analysis was carried out on a sample of 190 studies, only 20% of which were eyewitness identification studies like those in the present meta-analysis. The correlation of .48 was computed on a separate, rather heterogeneous, subsample of 18 facial identification studies that reported an accuracy–confidence correlation. Nevertheless, should this result be replicated, additional support for the optimality hypothesis will have been obtained. According to this hypothesis, the accuracy–confidence correlation should increase with increases in the opportunity to observe the target face. The latter increases certainly allow for more optimal information-processing circumstances.

Method

In the present review, a homogeneous sample of studies was taken with respect to the task expected of the witness, the measure of memory accuracy, and the manner of calculating the accuracy–confidence correlation (see Appendix for citations). A comprehensive search produced a sample of 35 studies that met the joint criteria of testing recognition memory for unfamiliar faces seen live only once during a staged presentation, of using postdecision confidence rating scales, and of computing interindividual accuracy–confidence correlations. This homogenization should have had the effect of reducing unnecessary variation in the size of r and should render generalization of the results less hazardous.

Two points should be made clear vis-à-vis procedures followed in cumulating results within and between studies. Following the advice of Hunter et al. (1982), the total group r for each study was entered into the initial meta-analysis. Where the total r was not given or could not be computed, then subgroup correlations were averaged and the average r was entered along with the total group sample size. As Hunter et al. (1982) have noted, this particular mean r will usually be slightly smaller than the appropriate total group r . Again, following Hunter et al.'s advice, averaging of r s within and between studies was accomplished without first converting raw r s to Fisher z' scores. Hunter et al. have shown that the usual procedure of averaging z' scores may produce an inflated estimate of mean r .²

Results

Across all 35 studies ($N = 3,953$) the weighted average $r = .2519$ ($r^2 = .0635$), whereas the equivalent $d = .5204$. The standard deviation of r corrected for sampling error was .0872 (.1261, uncorrected), yielding a 95% confidence interval of .08 to .42. Thus, the confidence interval excludes zero and also includes a value (.42) that, if it were the true value of r , would

¹ The interindividual accuracy–confidence correlation computed in eyewitness identification studies is arguably a more appropriate measure for generalization to the courtroom than is an intraindividual correlation. In the former case, a point-biserial correlation is computed between the dichotomous variable correct–incorrect on the single identification trial and rated confidence in the identification decision. Each of a large number of eyewitnesses contributes a pair of numbers to the correlation. In the latter instance, a correlation is computed for each eyewitness across multiple face recognition trials and multiple ratings of confidence; these within-subjects correlations can then be averaged. Such correlations would be of forensic relevance only if the court were interested in predicting a witness's true score accuracy at face recognition from a perfectly reliable measure of confidence. We rather suspect that the court is more interested in inferring the witness's accuracy at recognizing a singular suspect's face. Consequently, a better way to predict such accuracy would be to compute the accuracy–confidence correlation across large numbers of persons exposed to a single identification event. Averaging interindividual correlations across large numbers of separate suspects might then come closer to a proper estimate of the predictability of accuracy from confidence for any particular witness.

² For a sample size of 30 and a population correlation of .30, Silver and Dunlap (1987) have shown that the procedure for averaging raw r s underestimates the true r by about .002, and the procedure of converting to Fisher z' scores and then converting back overestimates by .004. Given the extra effort in executing the latter procedure, they see little response to adopt it for sample sizes greater than 30. Inasmuch as the present sample size is 35, we have adopted the former, more conservative procedure.

mean that as much as 18% of the variance in eyewitness accuracy could be accounted for by variations in witness confidence.

Inasmuch as sampling error is typically the largest source of statistical artifact in meta-analyses (Hunter et al., 1982), it was of interest to learn that sampling error accounted for 52.2% of the variation in r across these 35 studies. Even so, plenty of room remains not only for measurement error but also for moderator variables to produce observed variation in r .³ Unfortunately, because investigators have not supplied information concerning reliability of their accuracy and confidence measures, it was not possible to correct for measurement error. Of course, this means that the estimate of r is lower than it would otherwise have been.

The viability of several candidates for moderator variables was tested, as well. For reasons cited earlier, duration of target face exposure was tested first. The Pearson r calculated between study accuracy–confidence correlations and their respective target exposure durations was .356 ($r^2 = .1267$, $n = 34$, $p < .025$). It is possible, however, to correct this correlation for the attenuating effect of sampling error in the accuracy–confidence correlations. The meta-analysis revealed that 52.2% of the variance in those correlations is sampling error. This is the same as saying that 52.2% of the variance in measurement scale scores is random error of measurement. Hence, the reliability is $1 - .522 = .478$. Therefore, the correlation between the actual accuracy–confidence correlations (population values) and the exposure times is $.356/(\sqrt{.478}) = .5149$ ($r^2 = .2651$, $p < .005$). Note that the study correlations were not weighted by sample size in this instance because it made no sense to weight target exposure durations by sample size.

To gain an appreciation of how this variable might affect accuracy–confidence correlations that had been weighted by sample size, meta-analyses were performed on two subsamples defined by a median split of exposure interval ($C_{50} = 74.64$ s, $M = 68.56$ s, $s = 60.98$ s). Inasmuch as duration was not specified in one study and could not be reasonably estimated, there were 17 studies in each subsample. Estimated r in the shorter duration subsample ($n = 1,880$) was .1932 ($r^2 = .0373$), with $d = .3933$. The corrected standard deviation of r was .0964, yielding a 95% confidence interval of .00 to .38. For the longer duration subsample ($n = 1,980$), the weighted average r was .3091 ($r^2 = .0955$), whereas $d = .6497$. Given a corrected standard deviation of r of .08, the confidence interval extends from .15 to .47, an interval not including zero as does that for the shorter duration subsample.

Other variables that could be coded for were retention interval, lineup size, and whether or not a target-absent (TA) lineup was used, as well as a target-present (TP) one. None of the correlations between study accuracy–confidence correlations and any of these variables were significant. The r for retention interval was .2465 but shrank to .1403 when the study of Malpass and Devine (1981) was excluded—a study that used a memory enhancement manipulation that tended to counteract the effect of a 5-month retention interval. The weighted average accuracy–confidence correlation for 15 studies that included both TA and TP lineups was .2413, a value virtually identical to that for the entire sample of 35 studies. Finally, lineup size correlated .2675 with the accuracy–confidence correlation, a value rather close to that obtained by Shapiro and Penrod (1984),

who obtained an r of $-.23$ between the accuracy–confidence correlation and the ratio of targets to decoys at recognition test. This ratio becomes smaller as lineup size increases in eyewitness identification studies. So in both instances, a modest but not statistically reliable tendency was noted for larger recognition test arrays to be associated with a higher accuracy–confidence correlation.

Discussion

A perusal of the present findings prompts several conclusions. First, it would seem a conservative statement to conclude, on the basis of studies currently available, that the best estimate of the population correlation of eyewitness confidence and accuracy at recognizing an unfamiliar face is at least .25. The principal reason for .25 being a conservative estimate is that correction for the attenuating effects of measurement error could not be carried out. Had we been able to make this correction, the correlation would, of course, have been higher.

It is intriguing to speculate just how much higher the correlation would have been. The obtained r may well be closer to an unbiased estimate of the desired population correlation than one might think. The primary reason is that it may not be appropriate to correct the obtained correlation for unreliability in the accuracy measure. After all, the accuracy measure in the present sample does not represent the average level of accuracy across a large number of identification trials for the typical eyewitness (true score accuracy). Rather, a given witness is either correct or incorrect on a single identification trial, by its nature not a very reliable measurement. Correcting for unreliability in this measure would produce a spuriously high estimate of the interindividual correlation of accuracy and confidence (see Footnote 1).

It might be appropriate, however, to correct the obtained correlation for unreliability in the confidence measure, at least up to the level of reliability with which confidence would typically be assessed by jurors.⁴ The highest estimate of this reliability that we have seen is an r of .549 (Wells et al., 1979). Hence, the optimal estimate of the interindividual correlation of accuracy and confidence would be $(.2519)/(\sqrt{.549}) = .34$.

In any event, even a correlation of .25 cannot be characterized as an effect of negligible size. Inasmuch as Cohen (1977) has defined a small effect size as a d of .20, and a medium effect size as a d of .50, it would appear that the confidence–accuracy relation is medium size in nature, the overall estimated r of .25 being equivalent to a d of .52. This is in contrast to the negligible effect size reported by Wells and Murray (1984), a d of .14.

With the present meta-analysis we have been able, as well, to identify a modestly powerful moderator variable, *target exposure duration*. Thus, with a sample more homogeneous and twice as large, we have confirmed the result of Shapiro and Penrod (1984). Taken together, these findings are supportive of Def-

³ Range effects were probably not much a contributor to variation in r across the 35 studies sampled. If the subjects in each study were indeed random samples of the population of young adults enrolled in college, there should have been no artificial enhancement or suppression of our estimate of the true r due to this particular source of variance.

⁴ We are indebted to an anonymous reviewer for this suggestion.

fenbacher's (1980) optimality hypothesis, in that longer durations permit more efficient processing of faces to which eyewitnesses have been exposed. As the optimality hypothesis would predict, such efficiencies of processing allow for greater predictability of recognition memory from confidence. Apparently, as much as 27% of the variation in the predictability of accuracy from confidence can be accounted for by variation in the opportunity to observe a target person's face.

Even though the currently available studies show considerable variation in the amount of target exposure time, they do not show similar variation in level of eyewitness arousal, a variable we suspect may be a powerful moderator of how well confidence and accuracy are correlated in real-life witnessing conditions. Only one of the present sample of studies involved an explicit manipulation of arousal, Clifford and Hollin (1981). Of interest here is the fact that heightened arousal appeared to result in decreased predictability of recognition accuracy from confidence. In the Clifford and Hollin study, accuracy and confidence were correlated .41 ($p < .05$) for those witnesses who had viewed a nonviolent videotape. The correlation dropped to a value not significantly different from zero for witnesses who had viewed a violent tape. Supportive of this finding is a result obtained by Brigham et al. (1983). For the witnesses in the Brigham et al. study, within-subject correlations of accuracy and confidence were calculated. Whereas 50% of those in a condition of moderate arousal exhibited significant correlations, only 20% of those in a condition of high arousal did.

This evidence is obviously only suggestive. What it suggests is that arousal may be an important real-life moderator variable and that under conditions of high arousal, there may be little, if any, correlation of recognition memory and confidence. Unfortunately, research testing this proposition will be difficult to conduct. Witnesses would need to be exposed to realistically high levels of arousal in a situation in which their attributions would be appropriate to persons exposed to such levels.

There may well be other real-life sources of systematic distortion and random noise that serve to destroy much of the accuracy-confidence correlation in actual cases. We will mention just two: First, pretrial rehearsals with attorneys presumably bolster witness confidence but do nothing for accuracy, a case of systematic distortion. Second, additional random error could be introduced because measures of confidence in actual cases may be less precise than measures taken in laboratory or field studies. After all, in real-life situations, confidence is "mea-

sured" by subjective impressions of those listening to and observing the witness, whereas simulated jurors are provided with a metric for objectifying their judgments of witness confidence.

References

- Brigham, J. C., Maass, A., Martinez, D., & Whittenberger, G. (1983). The effect of arousal on facial recognition. *Basic and Applied Social Psychology*, 4, 279-293.
- Clifford, B. R., & Hollin, C. R. (1981). Effects of the type of incident and the number of perpetrators on eyewitness memory. *Journal of Applied Psychology*, 66, 364-370.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.
- Cutler, B. L., Penrod, S. D., & Stuve, T. E. (in press). Juror decisionmaking in eyewitness identification cases. *Law and Human Behavior*.
- Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relationship? *Law and Human Behavior*, 4, 243-260.
- Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin*, 88, 359-369.
- Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.
- Jenkins, F., & Davies, G. (1985). Contamination of facial memory through exposure to misleading composite pictures. *Journal of Applied Psychology*, 70, 164-176.
- Malpass, R. S., & Devine, P. G. (1981). Guided memory in eyewitness identification. *Journal of Applied Psychology*, 66, 343-350.
- Neil v. Biggers, 409 U.S. 188, 93 St. Ct. 375, 34 L. Ed. 2d 401 (1972).
- Pigott, M., & Brigham, J. C. (1985). Relationship between accuracy of prior description and facial recognition. *Journal of Applied Psychology*, 70, 547-555.
- Shapiro, P. N., & Penrod, S. D. (1984, August). *Meta-analysis of facial identification literature*. Paper presented at the meeting of the American Psychological Association, Toronto, Ontario, Canada.
- Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used? *Journal of Applied Psychology*, 72, 146-148.
- Wells, G. L. (1985). Verbal descriptions of faces from memory: Are they diagnostic of identification accuracy? *Journal of Applied Psychology*, 70, 619-626.
- Wells, G. L., & Hryciw, B. (1984). Memory for faces: Encoding and retrieval operations. *Memory & Cognition*, 12, 338-344.
- Wells, G. L., Lindsay, R. C. L., & Ferguson, T. J. (1979). Accuracy, confidence, and juror perceptions in eyewitness identification. *Journal of Applied Psychology*, 64, 440-448.
- Wells, G. L., & Murray, D. M. (1984). Eyewitness confidence. In G. L. Wells & E. F. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives* (pp. 155-170). New York: Cambridge University Press.

Appendix

References for the Meta-Analysis

- Bothwell, R. K., Brigham, J. C., & Pigott, M. A. (1987). *Personality, anxiety, and eyewitness memory*. Manuscript submitted for publication.
- Brigham, J. C., & Cairns, D. L. (1987). *The effect of mugshot inspections on eyewitness identification accuracy*. Manuscript submitted for publication.
- Brigham, J. C., Maass, A., Snyder, L. D., & Spaulding, K. (1982). Accuracy of eyewitness identifications in a field setting. *Journal of Personality and Social Psychology*, 42, 673-681.
- Brigham, J. C., & Pigott, M. A. (1983). *Comparison of description and identification accuracy under different attentional levels*. Unpublished manuscript. Florida State University.
- Clifford, B. R., & Hollin, C. R. (1981). Effects of the type of incident and number of perpetrators on eyewitness testimony. *Journal of Applied Psychology*, 66, 364-370.
- Cutler, B. L., & Penrod, S. D. (1987). *Improving the reliability of eyewitness identifications: Lineup construction and presentation*. Manuscript submitted for publication.
- Cutler, B. L., Penrod, S. D., & Martens, T. K. (1987). The reliability of eyewitness identification: The role of system and estimator variables. *Law and Human Behavior*, 11, 233-258.
- Cutler, B. L., Penrod, S. D., & Martens, T. K. (in press). Improving the reliability of eyewitness identifications: Putting context into context. *Journal of Applied Psychology*.
- Cutler, B. L., Penrod, S. D., O'Rourke, T. E., & Martens, T. K. (1986). Unconfounding the effects of contextual cues on eyewitness identification accuracy (Experiments 1 and 2). *Social Behavior*, 1, 113-134.
- Fleet, M. L., Brigham, J. C., & Bothwell, R. K. (1987). The confidence-accuracy relationship: The effects of confidence assessment and choosing. *Journal of Applied Social Psychology*, 17, 171-187.
- Gorenstein, G. W., & Ellsworth, P. C. (1980). Effect of choosing an incorrect photograph on a later identification by an eyewitness. *Journal of Applied Psychology*, 65, 616-622.
- Greenberg, M. S., Wilson, C. E., Ruback, R. B., & Mills, M. K. (1979). Social and emotional determinants of victim crime reporting. *Social Psychology Quarterly*, 42, 364-372.
- Hilgendorf, E. L., & Irving, B. L. (1978). False positive identification. *Medicine, Science and the Law*, 18, 255-262.
- Hosch, H. M., & Cooper, D. S. (1982). Victimization as a determinant of eyewitness accuracy. *Journal of Applied Psychology*, 67, 649-652.
- Hosch, H. M., Leippe, M. R., Marchioni, P. M., & Cooper, D. S. (1984). Victimization, self-monitoring, and eyewitness identification. *Journal of Applied Psychology*, 69, 280-288.
- Jenkins, F., & Davies, G. (1985). Contamination of facial memory through exposure to misleading composite pictures. *Journal of Applied Psychology*, 70, 164-176.
- Kassin, S. M. (1985). Eyewitness identification: Retrospective self-awareness and the accuracy-confidence correlation. *Journal of Personality and Social Psychology*, 49, 878-893.
- Krafka, C., & Penrod, S. (1985). Reinstatement of context in a field experiment on eyewitness identification. *Journal of Personality and Social Psychology*, 49, 58-69.
- Leippe, M. R., Wells, G. L., & Ostrom, T. M. (1978). Crime seriousness as a determinant of accuracy in eyewitness identification. *Journal of Applied Psychology*, 63, 345-351.
- Lindsay, R. C. L., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology*, 70, 556-564.
- Lindsay, R. C. L., Wells, G. L., & Rumpel, C. M. (1981). Can people detect eyewitness-identification accuracy within and across situations? *Journal of Applied Psychology*, 66, 79-89.
- Malpass, R. S., & Devine, P. G. (1981a). Eyewitness identification: Lineup instructions and the absence of the offender. *Journal of Applied Psychology*, 66, 482-489.
- Malpass, R. S., & Devine, P. G. (1981b). Guided memory in eyewitness identification. *Journal of Applied Psychology*, 66, 343-350.
- Murray, D. M., & Wells, G. L. (1982). Does knowledge that a crime was staged affect eyewitness accuracy? *Journal of Applied Social Psychology*, 12, 42-53.
- Pigott, M., & Brigham, J. C. (1985). Relationship between accuracy of prior description and facial recognition. *Journal of Applied Psychology*, 70, 547-555.
- Pigott, M. A., Brigham, J. C., & Bothwell, R. K. (1987). *A field study on the relationship between quality of eyewitnesses' descriptions and identification accuracy*. Manuscript submitted for publication, Florida State University.
- Platz, S. J., & Hosch, H. M. (in press). Cross-racial/ethnic eyewitness identification: A field study. *Journal of Applied Social Psychology*.
- Sanders, G. S., & Warnick, D. (1980). Some conditions maximizing eyewitness accuracy: A learning/memory analogy. *Journal of Criminal Justice*, 8, 395-403.
- Sanders, G. S., & Warnick, D. H. (1981). Truth and consequences: The effect of responsibility on eyewitness behavior. *Basic and Applied Social Psychology*, 2, 67-79.
- Shepherd, J. W., Ellis, H. D., & Davies, G. M. (1982). *Identification evidence: A psychological evaluation*. Aberdeen, Scotland: Aberdeen University Press.
- Wells, G. L., Ferguson, T. J., & Lindsay, R. C. L. (1981). The tractability of eyewitness confidence and its implications for triers of fact. *Journal of Applied Psychology*, 66, 688-696.
- Wells, G. L., & Leippe, M. R. (1981). How do triers of fact infer the accuracy of eyewitness identifications? Using memory for peripheral detail can be misleading. *Journal of Applied Psychology*, 66, 682-687.
- Wells, G. L., Lindsay, R. C. L., & Ferguson, T. J. (1979). Accuracy, confidence, and juror perceptions in eyewitness identification. *Journal of Applied Psychology*, 64, 440-448.
- Yuille, J. C., & McEwan, N. H. (1985). Use of hypnosis as an aid to eyewitness memory. *Journal of Applied Psychology*, 70, 389-400.

Received May 5, 1986

Revision received May 18, 1987

Accepted April 18, 1987 ■

SHORT NOTE

Predicting Supervisory Ratings Versus Promotional Progress in Test Validation Studies

Herbert H. Meyer
University of South Florida

Two test validation studies are described in which performance-over-time (promotional progress) yielded significantly higher correlations with ability test predictors than did supervisory ratings of present performance. The point is made that although supervisor ratings are used as the performance criterion in the great majority of validation studies, performance-over-time, as represented in promotional progress, may often provide a more reliable and valid criterion measure of job performance.

Most criterion-related validation studies use as the dependent variable a measure of job performance at a particular point in time. In the great majority of cases, this performance measure consists of supervisory ratings. Yet, in many validation studies a longer range criterion of performance over time, such as promotional progress, might also be available to the investigator. A number of studies of the validity of assessment centers in predicting managerial performance, for example, have used subsequent promotional progress as performance criteria (Thornton & Byham, 1982). In most cases, such a longer range criterion would seem to be more objective and more stable than supervisor ratings of performance. Nevertheless, in discussing research on the validity of assessment centers, Thornton and Byham (1982) stated,

The predominant criterion for the validation research in this body of literature has been supervisory *ratings* of both performance and potential. Virtually every study has used ratings, even when more objective data were also gathered, such as salary progress. . . . [They then went on to say,] Problems with supervisors' ratings are legion. Leniency, halo, and restrictions-in-range biases may occur (p. 298).

Recently, the author had two opportunities to compare the predictability of supervisory ratings of performance at a point in time with measures of performance-over-time—specifically, salary-for-age in one study and time-to-promotion out of trainee status in a second study.

Study 1

In Study 1, a small, technically oriented company had used the Wesman Personnel Classification Test (Wesman, 1965) for many years in screening new college-graduate job applicants. However, for about 10 years before this follow-up validation study was conducted, the company discontinued using the test scores in the screening process because of pressure from the Equal Employment Opportunity Commission (EEOC). Nevertheless, they continued to test White applicants, even though the scores were allegedly ignored. Test scores were not revealed

to supervisors and did not influence starting salaries in any way. Minority applicants were not tested, inasmuch as the test did prove to have significant adverse impact, and company management wanted to avoid any potential EEOC action against them.

Test scores were available in the files for 141 employees who had been tested after the use of scores in the screening process had allegedly been discontinued. This meant that the range of scores for that group was relatively unrestricted. Actually, a comparison of average total scores on the test for those hired and those rejected revealed that the scores may have been used to some extent in making selection decisions. The difference in favor of those selected was statistically significant ($t = 2.59$). However, it was not great from a practical standpoint—about 2.5 points on distributions with means of about 43 and standard deviations of about 8.

Performance Criteria

Two measures of job performance success achieved were used in the data analysis. The first measure was a current performance rating made by supervisors for salary administration purposes. These ratings were made, allegedly with great care, on a 4-point scale, with 1 the highest rating, and 4 the lowest. The distribution of these ratings was as follows: *high* (1), 9%; (2), 34%; (3), 35%; and *low* (4), 22%. The second measure, a career progress index, consisted of salary-for-age, inasmuch as all subjects had been hired at about the same age, as new college graduates. This index was a standard score on a scale of 1 to 9, which indicated the extent to which each person's salary deviated from the average salary being earned (5 on the scale) by persons in his or her age group. An inspection of salary distributions showed that after 2 years, top earners were making about 20% more than bottom earners. However, after 8–10 years, those in the top decile in earnings were being paid salaries that averaged more than twice the amount being paid to those in the bottom decile.

Validity Coefficients

Product-moment correlation coefficients computed between total scores on the Personnel Classification Test and the two performance criteria were .09 for the supervisory ratings of current job performance and .42 for the salary-for-age or performance-over-time indexes. The first correlation is obviously not significant, being only about .10, even if corrected for the fact that the supervisory ratings were grouped into only four categories (Pearson, 1913). The second correlation (.42), on the other hand, is significant beyond the .001 level. The difference between the two correlations is also significant well beyond the .001 level.

Study 2

In Study 2, a company with extensive computer operations had been using the SRA Computer Programmer Aptitude Battery (SRA, 1967) for more than 3 years to select computer operators. Although the test proved to have adverse impact for minority applicants, this effect had been minimized by using lower cutting scores for minorities. Nevertheless, EEOC representatives challenged the company to prove the validity of the test for selecting operators, inasmuch as the test had been designed to select programmers.

Test scores were available in employment files for 105 present operators who had started as trainees. Of the 105 employees, 47 were classified as minorities—30 Hispanics and 17 Blacks.

Performance Criteria

Two performance criterion measures were obtained for each operator: (a) supervisory ratings of the performance of the person as a trainee, and (b) time served in the trainee status. The latter index varied from 5 to 21 months, with a mean of about 9 months. The supervisor-instructors indicated that mastery of the technical aspects of the job was the primary criterion used in deciding when a trainee should be promoted to operator status.

Validity Coefficients

Correlations computed between total scores on the test and the two performance criteria were .11 for the supervisory ratings and .40 for the time-to-promotion index. As was the case in the first study, the correlation with supervisory ratings was not statistically significant, whereas the correlation with the time-to-promotion index was significant beyond the .001 level. Moreover, this correlation of .40 was *not* corrected for restriction of range, even though the cutting scores used eliminated well over one half of the applicants tested. The difference between the two correlations was significant at the .02 level.

Discussion

In both of the validation studies described, the longer range, promotional progress performance criterion correlated significantly with test scores, whereas the criterion based on supervisory ratings did not. In both studies, the ratings were carefully obtained and focused on ability to do the work—that is, the

aspect of job performance the tests were designed to predict. Separate ratings were obtained of motivational and attitudinal factors, in the hope that these considerations would be less likely to influence the supervisors in rating ability. An attempt was made to partial out the effects of the ratings of motivational and attitudinal factors on the correlations between test scores and ratings of ability to do the work, but because of the small size of the latter correlations, the effects of this statistical control were insignificant (raised each correlation about .01).

In most work situations, performance-over-time, as represented in promotional progress, may be a more reliable performance measure than supervisory ratings of performance. Promotional decisions are likely to be based on the collective judgments of a number of supervisors and managers. In many cases, objective performance data may be available to influence the promotion decisions. In the computer operator study, for example, trainees were given tests from time to time to ensure that they had mastered various technical aspects of the operator job. The supervisors must have used those performance indicators in making their promotional decisions, but they seem to have largely ignored them in making their ability ratings for this study.

Our findings in these studies are consistent with the findings of Schmitt, Gooding, Noe, and Kirsch (1984), who conducted a meta-analysis of 350 validation studies conducted between 1964 and 1982. They found that “objective” criteria, such as status change or earnings, yielded higher validity coefficients, on the average, than did performance ratings.

In many validation studies, both types of criteria may be available to the investigator. In some cases, a performance-over-time measure may be difficult to obtain and may be of questionable validity. In most cases, however, the performance-over-time measure will be easier to obtain than supervisory ratings; it merely requires a compilation and analysis of information available in files. Supervisor ratings available in the files, on the other hand, are rarely useful for research purposes, inasmuch as they are likely to be ridden with such errors as leniency and halo. Performance ratings useable for research often can only be obtained with a great deal of time, effort, and technical expertise. Why do we continue to rely so heavily on them in our validation research?

References

- Pearson, K. (1913). On the measurement of the influence of “broad categories” on correlation. *Biometrika*, 9, 116–139.
- Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology*, 37, 407–422.
- SRA (1967). *Computer Programmer Aptitude Battery*. Chicago: Science Research Associates.
- Thornton, G. C., III, & Byham, W. C. (1982). *Assessment centers and managerial performance*. New York: Academic Press.
- Wesman, A. G. (1965). *The Wesman Personnel Classification Test*. San Antonio, TX: Psychological Corporation.

Received December 17, 1986

Revision received April 20, 1987

Accepted April 8, 1987 ■

Acknowledgment

In addition to our regular Consulting Editors, we have used the services of many Ad Hoc Reviewers in the preparation of this volume of the Journal. We sincerely appreciate their contributions. Those reviewing for Volume 72 are the following:

Clifford Abe	Angelo S. DeNisi	Susan Jackson
Michael A. Abelson	Wayne R. Dexter	Rick R. Jacobs
Edward F. Adams	Robert L. Dipboye	P. Richard Jeanneret
Jerome Adams	Michael E. Doherty	Thomas H. Jerdee
Clayton P. Alderfer	Anthony N. Doob	Elaine M. Justice
John W. Aldrich, Jr.	Dennis L. Dossett	
Carol H. Ammons	George F. Dreher	Velma A. Kameoka
Nancy S. Anderson	Marvin D. Dunnette	Jeffrey S. Kane
William H. Angoff		Ruth Kanfer
John A. Antoinetti	Janet Echemendia	Rabindra N. Kanungo
Hugh J. Arnold	Dov Eden	Harold H. Kassarian
Richard D. Arvey	Jack Edwards	Edward S. Katkin
J. William Asher	Howard E. Egeth	Raymond A. Katzell
William G. Austin	Phoebe C. Ellsworth	Michael J. Kavanagh
	Miriam Erez	Jerard F. Kehoe
William K. Balzer	Martin G. Evans	David A. Kenny
Christina G. Banks		Steven Kerr
Gerald V. Barrett	Charles H. Fay	Charles A. Kiesler
Alan R. Bass	Gerald R. Ferris	Moses N. Kiggundu
Andrew S. Baum	Cynthia D. Fisher	Bill N. Kinder
Max H. Bazerman	Irene H. Frieze	David Kipnis
Chris J. Berger	Dean E. Frost	Richard J. Klimoski
Leonard Berkowitz	John J. Furedy	Dale Klopfer
Nicholas J. Beutell		Charles Y. Kondo
Vidya Bhushan	Martin Gannon	Richard E. Kopelman
Richard S. Blackburn	Joseph E. Garcia	Kurt Kraiger
Bruce Bloxom	Lewis R. Goldberg	Alan G. Kraut
Virginia R. Boehm	James G. Goodale	Allen I. Kraut
John W. Boudreau	Reginald A. Goodfellow	S. David Kriska
Michael T. Brannick	Michael E. Gordon	
Arthur P. Brief	Gary D. Gottfredson	David Lachar
Kelly J. Brookhouse	Harrison G. Gough	Charles E. Lance
Ronald J. Burke	George B. Graen	Darryl A. Lang
Robert F. Burnaska	Charles N. Greene	John M. Larsen, Jr.
	Robert A. Gregory	Kenneth R. Laughery
David E. Campbell	Dennis M. Groner	Robert L. Linn
James E. Campion	Paul V. Gump	Elizabeth F. Loftus
Richard Carlson	Rhonda L. Gutenberg	Manuel London
John C. Cavanaugh	Nina Gupta	Robert G. Lord
Gregory W. Cermak	Richard A. Guzzo	
John K. Chadwick-Jones		Robert MacCallum
Norman Cliff	Rick D. Hackett	Joseph M. Madden
Samuel W. Cochran	June Hahn	Charles A. Maher
Andrew L. Comrey	Paul J. Hanges	Roy S. Malpass
Robert L. Conner	Lenore W. Harmon	Melvin M. Mark
Paul Cook	David Harrison	Christina Maslach
William H. Cooper	David M. Herold	John Mathieu
John W. Cotton	Harmon M. Hosch	Jerry R. May
Lee J. Cronbach	Yi-Ming Hsu	George Milkovich
Sara J. Czaja	Charles L. Hulin	Alan N. Miller
	Lloyd G. Humphreys	Norman Miller
Fred E. Dansereau, Jr.	Joe Hurrell	Shitala P. Mishra
Graham M. Davies		Robert F. Morrison
René V. Dawis	John M. Ivancevich	Rudolf G. Mortimer
Michael E. Dawson	Chizuko Izawa	Stephan J. Motowidlo
Tom DeCotiis		Paul M. Muchinsky
James E. Deese		Kevin R. Murphy

Lawrence R. Murphy
Paul Muter

Kevin McConkey
Joseph E. McGrath
D. Douglas McKenna
Mary McLaughlin

Franco M. Nicosia
Walter R. Nord
Kent L. Norman
Warren T. Norman

Frank J. Ofsanko
Brian S. O'Leary
Terence A. Oliva
Engram Olkin
Judith Olson
Charles A. O'Reilly, III.
Dennis W. Organ
Hobart G. Osburn

Michael S. Pallack
Kenneth L. Pargament
Kenneth Pearlman
Steven Penrod
Erich P. Prien
Christopher Puto

Mohammed Y. Quereshi

Nambury S. Raju
Robert A. Ramos
Bikkar S. Randhawa

John M. Rauschenberger
Richard R. Reilly
Peter H. Reingen
Deborah R. Richardson
Benson Rosen
Richard M. Rozelle
Craig J. Russell
Sara L. Rynes

Lisa M. Saari
Marshall Sashkin
Marvin G. Scherr
Benjamin Schneider
Ellin K. Scholnick
John R. Schuck
Donald P. Schwab
Clive R. Seligman
Saul B. Sells
Richard J. Shavelson

Samuel Shye
Barry Smith
Carlla S. Smith
Melvin Sorcher
Wendy G. Soubel
Paul E. Spector
Janet T. Spence
Mark D. Spool
Ross Stagner
Susan W. Stang
David W. Stewart
Stanley Stephenson
Robert M. Stern
Eugene F. Stone
Mitzi Svoboda

John A. Swets
Fred S. Switzer, III

Lois E. Tetrick
Julian Thayer
Paul W. Thayer
Howard E. A. Tinsley
John Tisak
Michael A. Todt
Jay Tombaugh
Molly Treadway
Oliver C. S. Tzeng
Aharon Tziner

Enzo R. Valenzi
Robert J. Vance
Philip E. Varca
Nu Viet Vu

William M. Waid
Lawrence K. Waters
Howard M. Weiss
Peter Weissenberg
Kenneth N. Wexley
Richard A. Winett
Myron Wolbarsht
Gerrit Wolf
William Wooten

Kim O. Yap
Kenneth M. York

Stephen J. Zaccaro
Mary D. Zalesny

U.S. Postal Service STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION (Required by 39 U.S.C. 3685)		
1. TITLE OF PUBLICATION Journal of Applied Psychology		1B. PUBLICATION NO. 278360
2. DATE OF FILING 10/1/87		
3. FREQUENCY OF ISSUE Quarterly	4. COMPLETE MAILING ADDRESS OF THE PUBLISHER, EDITOR, AND MANAGING EDITOR (This item MUST NOT be omitted)	5. ANNUAL SUBSCRIPTION PRICE \$30member \$40indv-\$82inst
6. COMPLETE MAILING ADDRESS OF THE PUBLISHER, EDITOR, AND MANAGING EDITOR (This item MUST NOT be omitted)		
1400 N Uhle Street - Arlington, VA 22201		
7. COMPLETE MAILING ADDRESS OF THE PUBLISHER, EDITOR, AND MANAGING EDITOR (This item MUST NOT be omitted)		
1200 17th Street, NW, Washington, DC 20036		
8. FULL NAMES AND COMPLETE MAILING ADDRESSES OF PUBLISHER, EDITOR, AND MANAGING EDITOR (This item MUST NOT be omitted)		
PUBLISHER (Name and Complete Mailing Address) American Psychological Association - 1200 17th Street, NW-DC 20036		
EDITOR (Name and Complete Mailing Address) Robert Guion, Dept. of Psychol., Bowling Green State University, Bowling Green, OH 43403		
MANAGING EDITOR (Name and Complete Mailing Address) Susan Knapp - 1400 N Uhle Street - Arlington, VA 22201		
9. OWNER (If owned by a corporation, its name and address must be stated and also immediately thereunder the names and addresses of stockholders owning or holding 1 percent or more of total amount of stock. If not owned by a corporation, the names and addresses of the individual owners must be given. If owned by a partnership or other unincorporated firm, its name and address as well as that of each individual must be given. If the publication is published by a nonprofit organization, its name and address must be stated. (This item must be completed.)		
FULL NAME American Psychological Association		
COMPLETE MAILING ADDRESS 1200 17th Street NW Washington, DC 20036		
10. KNOWN BONDHOLDERS, MORTGAGEES, AND OTHER SECURITY HOLDERS (If none, so state)		
FULL NAME NONE		
COMPLETE MAILING ADDRESS		
11. FOR COMPLETION BY NONPROFIT ORGANIZATIONS AUTHORIZED TO MAIL AT SPECIAL RATES (Section 411.1) ONLY (The organization must be a nonprofit organization of the United States or a nonprofit organization of a foreign country. (Check one))		
(1) HAS NOT CHANGED DURING PRECEDING 12 MONTHS <input checked="" type="checkbox"/> (2) HAS CHANGED DURING PRECEDING 12 MONTHS <input type="checkbox"/> (If changed, publisher must submit explanation of change with this statement.)		
12. EXTENT AND NATURE OF CIRCULATION (See instructions on reverse side)		
A. TOTAL NO. COPIES (Net Press Run)		7,542
B. PAID AND/OR REQUESTED CIRCULATION		2,585
1. Sales through dealers and carriers, street vendors and counter sales		3,189
2. Mail Subscription (Paid and/or requested)		2,880
C. TOTAL PAID AND/OR REQUESTED CIRCULATION (Sum of 10B1 and 10B2)		5,465
D. DISTRIBUTION BY MAIL, CARRIER OR OTHER SAMPLES, COMPLIMENTARY, AND OTHER FREE COPIES		123
E. TOTAL DISTRIBUTION (Sum of C and D)		5,588
F. COPIES NOT DISTRIBUTED		2,428
1. Office use, left over, unaccounted, spoiled after printing		--
2. Return from Agents		--
G. TOTAL (Sum of E, F1 and 2 - should equal net press run shown in A)		8,016
13. I certify that the statements made by me above are true and complete		SIGNATURE AND TITLE OF EDITOR, PUBLISHER, BUSINESS MANAGER, OR OWNER Pat H. Mangano C.H. Sec Director

Author Index to Volume 72

Key to Pagination					
<i>Issue No.</i>	<i>Month</i>	<i>Pages</i>	<i>Issue No.</i>	<i>Month</i>	<i>Pages</i>
1	February	1-184	3	August	337-512
2	May	185-336	4	November	513-706

ARTICLES

Abelson, Michael A.—Examination of Avoidable and Unavoidable Turnover	382
Adelman, Pamela K.—Occupational Complexity, Control, and Personal Income: Their Relation to Psychological Well-Being in Men and Women	529
Alexander, Ralph A., and Barrick, Murray R.—Estimating the Standard Error of Projected Dollar Gains in Utility Analysis	475
Altmaier, Elizabeth— <i>see</i> Russell, Daniel W.	
Alvares, Kenneth M.— <i>see</i> Bycio, Peter	
Athey, Timothy R., and McIntyre, Robert M.—Effect of Rater Training on Rater Accuracy: Levels-of-Processing Theory and Social Facilitation Theory Perspectives	567
Barling, Julian, Bluen, Stephen D., and Fain, Rolene—Psychological Functioning Following a Disaster	683
Baron, Robert A.—Effects of Negative Ions on Cognitive Performance	131
Barrick, Murray R.— <i>see</i> Alexander, Ralph A.	
Beach, Rebecca— <i>see</i> Greenwald, Anthony G.	
Bennett, Corwin— <i>see</i> Silver, Edward M.	
Bentson, Cynthia— <i>see</i> Gaugler, Barbara B.	
Berkowitz, Leonard, Fraser, Colin, Treasure, F. Peter, and Cochran, Susan—Pay, Equity, Job Gratifications, and Comparisons in Pay Satisfaction	544
Blencoe, Allyn G.— <i>see</i> DeNisi, Angelo S.	
Bluen, Stephen D.— <i>see</i> Barling, Julian	
Bothwell, Robert K., Deffenbacher, Kenneth A., and Brigham, John C.—Correlation of Eyewitness Accuracy and Confidence: Optimality Hypothesis Revisited	691
Botzger, Preston C., and Yetton, Philip W.—Improving Group Performance by Training in Individual Problem Solving ..	651
Brigham, John C.— <i>see</i> Bothwell, Robert K.	
Bruning, Nealia S., and Frew, David R.—Effects of Exercise, Relaxation, and Management Skills Training on Physiological Stress Indicators: A Field Experiment	515
Bycio, Peter, Alvares, Kenneth M., and Hahn, June—Situational Specificity In Assessment Center Ratings: A Confirmatory Factor Analysis	463
Carnot, Catherine G.— <i>see</i> Greenwald, Anthony G.	
Carpenter, Bruce N.— <i>see</i> Raza, Susan M.	
Carsten, Jeanne M., and Spector, Paul E.—Unemployment, Job Satisfaction, and Employee Turnover: A Meta-Analytic Test of the Muchinsky Model	374
Cascio, Wayne F.— <i>see</i> Greer, Olen L.	
Cochran, Susan— <i>see</i> Berkowitz, Leonard	
Colarelli, Stephen M., Dean, Roger A., and Konstans, Constantine—Comparative Effects of Personal and Situational Influences on Job Outcomes of New Professionals	558
Comish, Sara Elizabeth—Recognition of Facial Stimuli Following an Intervening Task Involving the Identi-kit	488
Comrey, Andrew L.— <i>see</i> Montag, I.	
Conley, Patrick R., and Sackett, Paul R.—Effects of Using High-Versus Low-Performing Job Incumbents as Sources of Job-Analysis Information	434

Conlon, Edward J., and Parks, Judi McLean—Information Requests in the Context of Escalation	344
Constans, Joseph I.— <i>see</i> Murphy, Kevin R.	
Cornelius, Edwin T., III— <i>see</i> DeNisi, Angelo S.	
Cronshaw, Steven F., and Lord, Robert G.—Effects of Categorization, Attribution, and Encoding Processes on Leadership Perceptions	97
Cutler, Brian L., Penrod, Steven D., and Martens, Todd K.—Improving the Reliability of Eyewitness Identification: Putting Context Into Context	629
Dalton, Dan R., Todor, William D., and Owen, Crystal L.—Sex Effects in Workplace Justice Outcomes: A Field Assessment	156
Dean, Roger A.— <i>see</i> Colarelli, Stephen M.	
Deffenbacher, Kenneth A.— <i>see</i> Bothwell, Robert K.	
DeNisi, Angelo S., Cornelius, Edwin T., III, and Blencoe, Allyn G.—Further Investigation of Common Knowledge Effects on Job Analysis Ratings	262
de Wetter, Robert— <i>see</i> Mehrabian, Albert	
Dickinson, Terry L.— <i>see</i> Hamilton, John W.	
Drasgow, Fritz—Study of the Measurement Bias of Two Standardized Psychological Tests	19
Drasgow, Fritz, and Guertler, Elaine—A Decision-Theoretic Approach to the Use of Appropriateness Measurement for Detecting Invalid Test and Scale Scores	10
Drasgow, Fritz— <i>see</i> Idaszak, Jacqueline R.	
Dunlap, William P.— <i>see</i> Silver, N. Clayton	
Earley, P. Christopher, Wojnaroski, Pauline, and Prest, William—Task Planning and Energy Expended: Exploration of How Goals Influence Performance	107
Earley, P. Christopher— <i>see</i> Erez, Miriam	
Ellsworth, Phoebe C.— <i>see</i> Smith, Vicki L.	
Erez, Miriam, and Earley, P. Christopher—Comparative Analysis of Goal-Setting Strategies Across Cultures	658
Fain, Rolene— <i>see</i> Barling, Julian	
Fleishman, Edwin A.— <i>see</i> Mumford, Michael D.	
Ford, Robert C.— <i>see</i> McGee, Gail W.	
Fraser, Colin— <i>see</i> Berkowitz, Leonard	
Frayne, Colette A., and Latham, Gary P.—Application of Social Learning Theory to Employee Self-Management of Attendance	387
Frew, David R.— <i>see</i> Bruning, Nealia S.	
Fried, Yitzhak— <i>see</i> Oldham, Greg R.	
Gaugler, Barbara B., Rosenthal, Douglas B., Thornton, George C., III, and Bentson, Cynthia—Meta-Analysis of Assessment Center Validity	493
Gerhart, Barry—How Important Are Dispositional Factors as Determinants of Job Satisfaction? Implications for Job Design and Other Personnel Programs	366
Geyer, Paul D.— <i>see</i> Pond, Samuel B., III	
Gier, Joseph A.— <i>see</i> Weekley, Jeff A.	
Givon, Moshe M., and Goldman, Arie—Perceptual and Preferential Discrimination Abilities in Taste Tests	301
Gleason, Sandra E.— <i>see</i> Schmitt, Neal	

- Goldman, Arieh—*see* Givon, Moshe M.
- Greenberg, Jerald—Reactions to Procedural Injustice in Payment Distributions: Do the Means Justify the Ends? 55
- Greenwald, Anthony G., Carnot, Catherine G., Beach, Rebecca, and Young, Barbara—Increasing Voting Behavior by Asking People if They Expect to Vote 315
- Greer, Olen L., and Cascio, Wayne F.—Is Cost Accounting the Answer? Comparison of Two Behaviorally Based Methods for Estimating the Standard Deviation of Job Performance in Dollars With a Cost-Accounting-Based Approach 588
- Gross, Alan L., and McGanney, Mary Lou—The Restriction of Range Problem and Nonignorable Selection Processes 604
- Guastello, Stephen J.—A Butterfly Catastrophe Model of Motivation in Organizations: Academic Performance 165
- Guertler, Elaine—*see* Drasgow, Fritz
- Gutek, Barbara A.—*see* Schlenker, Judith A.
- Gutek, Barbara A.—*see* Schriber, Jacquelyn B.
- Hahn, June—*see* Bycio, Peter
- Hamilton, John W., and Dickinson, Terry L.—Comparison of Several Procedures for Generating J-Coefficients 49
- Hammer, Tove H., and Turk, Jay M.—Organizational Determinants of Leader Behavior and Authority 674
- Harding, Francis D.—*see* Mumford, Michael D.
- Heilman, Madeline E., Simon, Michael C., and Repper, David P.—Intentionally Favored, Unintentionally Harmed? Impact of Sex-Based Preferential Selection on Self-Perceptions and Self-Evaluations 62
- Helmreich, Robert L.—*see* Spence, Janet T.
- Henry, Rebecca A., and Hulin, Charles L.—Stability of Skilled Performance Across Time: Some Generalizations and Limitations on Utilities 457
- Hill, Thomas, Smith, Nancy D., and Mann, Millard F.—Role of Efficacy Expectations in Predicting the Decision to Use Advanced Technologies: The Case of Computers 307
- Hollenbeck, John R., and Klein, Howard J.—Goal Commitment and the Goal-Setting Process: Problems, Prospects, and Proposals for Future Research 212
- Hollenbeck, John R., and Williams, Charles R.—Goal Importance, Self-Focus, and the Goal-Setting Process 204
- Huber, Vandra L., and Neale, Margaret A.—Effects of Self- and Competitor Goals on Performance in an Interdependent Bargaining Task 197
- Hulin, Charles L.—*see* Henry, Rebecca A.
- Idaszak, Jacqueline R., and Drasgow, Fritz—A Revision of the Job Diagnostic Survey: Elimination of a Measurement Artifact 69
- Ilgen, Daniel R., and Moore, Carol F.—Types and Choices of Performance Feedback 401
- Josephs, Susan L.—*see* Latack, Janina C.
- Kabanoff, Boris—Predictive Validity of the MODE Conflict Instrument 160
- Kakuyama, Takashi—*see* Matsui, Tamao
- Kemery, Edward R., Mossholder, Kevin W., and Roth, Lawrence—The Power of the Schmidt and Hunter Additive Model of Validity Generalization 30
- Kirsch, Michael P.—*see* Kozlowski, Steve W. J.
- Klein, Howard J.—*see* Hollenbeck, John R.
- Klein, Katherine J.—Employee Stock Ownership and Employee Attitudes: A Test of Three Models 319
- Konstans, Constantine—*see* Colarelli, Stephen M.
- Kozlowski, Steve W. J., and Kirsch, Michael P.—The Systematic Distortion Hypothesis, Halo, and Accuracy: An Individual-Level Analysis 252
- Lane, Irving M.—*see* Prestholdt, Perry H.
- LaRocco, James M.—*see* Tetrick, Lois E.
- Larsson, Gerry—Routinization of Mental Training in Organizations: Effects on Performance and Well-Being 88
- Latack, Janina C., Josephs, Susan L., Roach, Bonnie L., and Levine, Mitchell D.—Carpenter Apprentices: Comparison of Career Transitions for Men and Women 393
- Latham, Gary P.—*see* Frayne, Colette A.
- Lautenschlager, Gary J., and Shaffer, Garnett Stokes—Reexamining the Component Stability of Owens's Biographical Questionnaire 149
- Leana, Carrie R.—Power Relinquishment Versus Power Sharing: Theoretical Clarification and Empirical Comparison of Delegation and Participation 228
- Levine, Edward L.—*see* Spector, Paul E.
- Levine, Mitchell D.—*see* Latack, Janina C.
- Libby, Robert, Trotman, Ken T., and Zimmer, Ian—Member Variation, Recognition of Expertise, and Group Performance 81
- Locke, Edwin A.—*see* Wood, Robert E.
- Lord, Robert G.—*see* Cronshaw, Steven F.
- Mann, Millard F.—*see* Hill, Thomas
- Marcus, Philip M.—*see* Schmitt, Neal
- Martens, Todd K.—*see* Cutler, Brian L.
- Mathews, Robert C.—*see* Prestholdt, Perry H.
- Matsui, Tamao, Kakuyama, Takashi, and Onglatco, Mary Lou Uy—Effects of Goals and Feedback on Performance in Groups 407
- McGanney, Mary Lou—*see* Gross, Alan L.
- McGee, Gail W., and Ford, Robert C.—Two (or More?) Dimensions of Organizational Commitment: Reexamination of the Affective and Continuance Commitment Scales 638
- McIntyre, Robert M.—*see* Athey, Timothy R.
- Meglino, Bruce M.—*see* Ravlin, Elizabeth C.
- Mehrabian, Albert, and de Wetter, Robert—Experimental Test of an Emotion-Based Approach to Fitting Brand Names to Products 125
- Mento, Anthony J.—*see* Wood, Robert E.
- Meyer, Herbert H.—Predicting Supervisory Ratings Versus Promotional Progress in Test Validation Studies 696
- Montag, I., and Comrey, Andrew L.—Internality and Externality as Correlates of Involvement in Fatal Driving Accidents 339
- Moore, Carol F.—*see* Ilgen, Daniel R.
- Mossholder, Kevin W.—*see* Kemery, Edward R.
- Motowidlo, Stephan J.—*see* Srinivas, Shanthi
- Mount, Michael K., and Thompson, Duane E.—Cognitive Categorization and Quality of Performance Ratings 240
- Mumford, Michael D., Weeks, Joseph L., Harding, Francis D., and Fleishman, Edwin A.—Measuring Occupational Difficulty: A Construct Validation Against Training Criteria 578
- Murphy, Kevin R.—Detecting Infrequent Deception 611
- Murphy, Kevin R., and Constans, Joseph I.—Behavioral Anchors as a Source of Bias in Rating 573
- Neale, Margaret A.—*see* Huber, Vandra L.
- Notz, William W., and Starke, Frederick A.—Arbitration and Distributive Justice: Equity or Equality? 359
- Oldham, Greg R., and Fried, Yitzhak—Employee Reactions to Workspace Characteristics 75
- Onglatco, Mary Lou Uy—*see* Matsui, Tamao
- Owen, Crystal L.—*see* Dalton, Dan R.
- Parkes, Katharine R.—Relative Weight, Smoking, and Mental Health as Predictors of Sickness and Absence From Work 275
- Parks, Judi McLean—*see* Conlon, Edward J.

- Penrod, Steven D.—*see* Cutler, Brian L.
- Pigozzi, Bruce—*see* Schmitt, Neal
- Pond, Samuel B., III, and Geyer, Paul D.—Employee Age as a Moderator of the Relation Between Perceived Work Alternatives and Job Satisfaction 522
- Pred, Robert S.—*see* Spence, Janet T.
- Prest, William—*see* Earley, P. Christopher
- Prestholdt, Perry H., Lane, Irving M., and Mathews, Robert C.—Nurse Turnover as Reasoned Action: Development of a Process Model 221
- Pryor, Robert G. L.—Differences Among Differences: In Search of General Work Preference Dimensions 426
- Puffer, Sheila M.—Prosocial Behavior, Noncompliant Behavior, and Work Performance Among Commission Salespeople 615
- Ravlin, Elizabeth C., and Meglino, Bruce M.—Effect of Values on Perception and Decision Making: A Study of Alternative Work Values Measures 666
- Raza, Susan M., and Carpenter, Bruce N.—A Model of Hiring Decisions in Real Employment Interviews 596
- Repper, David P.—*see* Heilman, Madeline E.
- Roach, Bonnie L.—*see* Latack, Janina C.
- Rosenthal, Douglas B.—*see* Gaugler, Barbara B.
- Roth, Lawrence—*see* Kemery, Edward R.
- Roznowski, Mary—Use of Tests Manifesting Sex Differences as Measures of Intelligence: Implications for Measurement Bias 480
- Rude, Dale E.—*see* Wall, James A., Jr.
- Russell, Daniel W., Altmaier, Elizabeth, and Van Velzen, Dawn—Job-Related Stress, Social Support, and Burnout Among Classroom Teachers 269
- Sackett, Paul R.—*see* Conley, Patrick R.
- Salthouse, Timothy A., and Saults, J. Scott—Multiple Spans in Transcription Typing 187
- Saults, J. Scott—*see* Salthouse, Timothy A.
- Schlenker, Judith A., and Gutek, Barbara A.—Effects of Role Loss on Work-Related Attitudes 287
- Schmitt, Neal, Gleason, Sandra E., Pigozzi, Bruce, and Marcus, Philip M.—Business Climate Attitudes and Company Relocation Decisions 622
- Schriber, Jacquelyn B., and Gutek, Barbara A.—Some Time Dimensions of Work: Measurement of an Underlying Aspect of Organization Culture 642
- Schurr, Paul H.—Effects of Gain and Loss Decision Frames on Risky Purchase Negotiations 351
- Severy, Lawrence J.—*see* Wilmoth, Gregory H.
- Shaffer, Garnett Stokes—Patterns of Work and Nonwork Satisfaction 115
- Shaffer, Garnett Stokes—*see* Lautenschlager, Gary J.
- Silver, Edward M., and Bennett, Corwin—Modification of the Minnesota Clerical Test to Predict Performance on Video Display Terminals 153
- Silver, N. Clayton, and Dunlap, William P.—Averaging Correlation Coefficients: Should Fisher's *z* Transformation Be Used? 146
- Silver, Starr—*see* Wilmoth, Gregory H.
- Simon, Michael C.—*see* Heilman, Madeline E.
- Smith, Nancy D.—*see* Hill, Thomas
- Smith, Vicki L., and Ellsworth, Phoebe C.—The Social Psychology of Eyewitness Accuracy: Misleading Questions and Communicator Expertise 294
- Spector, Paul E.—Method Variance as an Artifact in Self-Reported Affect and Perceptions at Work: Myth or Significant Problem? 438
- Spector, Paul E., and Levine, Edward L.—Meta-Analysis for Integrating Study Outcomes: A Monte Carlo Study of Its Susceptibility to Type I and Type II Errors 3
- Spector, Paul E.—*see* Carsten, Jeanne M.
- Spence, Janet T., Helmreich, Robert L., and Pred, Robert S.—Impatience Versus Achievement Strivings in the Type A Pattern: Differential Effects on Students' Health and Academic Achievement 522
- Srinivas, Shanthi, and Motowidlo, Stephan J.—Effects of Raters' Stress on the Dispersion and Favorability of Performance Ratings 247
- Starke, Frederick A.—*see* Notz, William W.
- Stone, Dianna L., and Stone, Eugene F.—Effects of Missing Application-Blank Information on Personnel Selection Decisions: Do Privacy Protection Strategies Bias the Outcome? 452
- Stone, Eugene F.—*see* Stone, Dianna L.
- Tetrick, Lois E., and LaRocco, James M.—Understanding, Prediction, and Control as Moderators of the Relationships Between Perceived Stress, Satisfaction, and Psychological Well-Being 538
- Thompson, Duane E.—*see* Mount, Michael K.
- Thornton, George C., III—*see* Gaugler, Barbara B.
- Todor, William D.—*see* Dalton, Dan R.
- Treasure, F. Peter—*see* Berkowitz, Leonard
- Trotman, Ken T.—*see* Libby, Robert
- Turk, Jay M.—*see* Hammer, Tove H.
- Van Velzen, Dawn—*see* Russell, Daniel W.
- Vecchio, Robert P.—Situational Leadership Theory: An Examination of a Prescriptive Theory 444
- Wall, James A., Jr., and Rude, Dale E.—Judges' Mediation of Settlement Negotiations 234
- Weekley, Jeff A., and Gier, Joseph A.—Reliability and Validity of the Situational Interview for a Sales Position 484
- Weeks, Joseph L.—*see* Mumford, Michael D.
- Williams, Charles R.—*see* Hollenbeck, John R.
- Wilmoth, Gregory H., Silver, Starr, and Severy, Lawrence J.—Receptivity and Planned Change: Community Attitudes and Deinstitutionalization 138
- Wojnarowski, Pauline—*see* Earley, P. Christopher
- Wood, Robert E., Mento, Anthony J., and Locke, Edwin A.—Task Complexity as a Moderator of Goal Effects: A Meta-Analysis 416
- Yetton, Philip W.—*see* Bottger, Preston C.
- Young, Barbara—*see* Greenwald, Anthony G.
- Zeidner, Moshe—Test of the Cultural Bias Hypothesis: Some Israeli Findings 38
- Zimmer, Ian—*see* Libby, Robert

OTHER

- Acknowledgment 698
- Call for Nominations for *Journal of Experimental Psychology: General* 528
- Call for Nominations for *Journal of Abnormal Psychology* 603
- Correction to Earley et al. 373
- Delworth Appointed Editor of *Professional Psychology: Research and Practice*, 1989–1994 566
- Instructions to Authors 137, 314, 492, 641
- Kintsch Appointed Editor of *Psychological Review*, 1989–1994 557
- Low Publication Prices for APA Members and Affiliates 54
- Schmitt Appointed Editor, 1989–1994 386, 551
- Searches Open for Editors of Five APA journals 124

SUPPORT FOR YOUR

RESEARCH

Find it with American Psychological Association's Guide to Research Support, Third Edition.

With competition for funding becoming increasingly intense, you *need* the vital advantage that the **Guide** will give you — whether this is your first proposal or the 100th!

The new, third edition of **APA's Guide to Research Support** provides *all the basic information behavioral scientists need to locate possible sources of support for their work*, covering over 180 federal programs and 70 foundations and other non profit organizations that support behavioral science research.

The **only** compilation of its type for the behavioral sciences, this new edition of the **Guide to Research Support** is a unique, comprehensive resource for the researcher. Its organization is streamlined to allow for greater ease of use.

Three time-saving indexes are provided.

- Program Name Index
- Subject Index
- Fellowship Sources

Each program description includes information on the types of research supported, proposed funding levels for 1987, as well as:

- Names, Telephone Numbers, and Addresses
- Funding Mechanisms
- Application Procedures
- Submission Deadlines
- Review Process
- Practical Advice

The **Guide** also provides data on the number, dollar range, average dollar amount, and duration of awards made in recent years.

This unique, comprehensive **Guide** is an *essential* reference for anyone seeking research support. If you need research funding, you need the **Guide to Research Support!** Order this timely source of funding information now!

YES! I would like to order this essential guide to research funding:

APA's Guide to Research Support, 3rd Edition

(Item Number 4250030)

_____ copies at \$22.00 (APA Members, Associates, and Affiliates)	_____
_____ copies at \$30.00 (List Price)	_____
Shipping & Handling	\$2.00
Total	_____

Orders totaling less than \$25.00 must be prepaid. Allow 4-6 weeks for delivery. Prices are subject to change without notice. UPS shipment is available at additional cost.

Charge my:

_____ VISA _____ Mastercard, No. _____

Exp. date _____

Signature _____

Ship to:

NAME _____

ADDRESS _____

CITY _____ STATE _____ ZIP _____

Make checks payable to APA and mail this order form to:

American Psychological Association

P.O. Box 2710

Hyattsville, MD 20784

To order by phone with Visa or Mastercard, call (703) 247-7705 (no collect calls, please).

DB

